

■

UNIVERSAL  
LIBRARY



127 616

UNIVERSAL  
LIBRARY



















■

PSYCHOLOGICAL TESTS  
METHODS, AND RESULTS

■



# A GROUP OF BOOKS IN PSYCHOLOGY

UNDER THE GENERAL EDITORSHIP OF

GARDNER MURPHY

*Columbia University*

---

## *GENERAL PSYCHOLOGY*

BY GARDNER MURPHY

## *PSYCHOLOGICAL TESTS, METHODS, AND RESULTS*

BY H. E. GARRETT AND M. R. SCHNECK

## *PSYCHOLOGY AND THE NEW EDUCATION*

BY S. L. PRESSEY



# PSYCHOLOGICAL TESTS, METHODS, AND RESULTS

BY

HENRY E. GARRETT, Ph.D.

ASSISTANT PROFESSOR OF PSYCHOLOGY  
COLUMBIA UNIVERSITY

AND

MATTHEW R. SCHNECK, Ph.D.

ASSOCIATE PROFESSOR OF PSYCHOLOGY  
UNIVERSITY OF ARIZONA



HARPER & BROTHERS, PUBLISHERS

NEW YORK AND LONDON

1933

■

PSYCHOLOGICAL TESTS, METHODS, AND RESULTS  
*Copyright, 1933, by Harper & Brothers*  
*Printed in the U. S. A.*

FIRST EDITION

F-11

*All rights in this book are reserved.*  
*No part of the text may be reproduced in any*  
*manner whatsoever without permission in writ-*  
*ing from Harper & Brothers.*

■

## PREFACE

THIS book is intended to serve both as a textbook and as a laboratory manual in courses in psychological tests. It is the hope of the authors that it may bridge the gap between the handbook of mental tests on the one hand, and the general textbook treatment on the other. The material in the book covers a wide range of topics, and it is possible for the instructor, who is giving a laboratory course, to select as many or as few of the tests as he sees fit.

Part One describes many tests not usually included in courses in mental testing, principally, we believe, because of the too great emphasis placed upon general intelligence tests. Tests of memory, learning, motor coordination and the like, however, besides being extremely useful to the applied psychologist, are really more *psychological*—in the sense of being measures of identifiable psychological functions—than are the batteries of paper-and-pencil tests designed to measure such complex functions as intelligence and personality. Part Two contains the tests which are probably most useful to students of education and of vocational psychology.

The chapter on statistical method usually found in books on tests has not been included for various reasons. Statistical methods have become so extensive that it is impossible in a single chapter to give the beginning student an adequate foundation. Moreover, we believe that the student, who intends to make himself really acquainted with psychological tests, should take a course in statistics before he begins the study of mental measurement.

The division of responsibility in the writing of this book should, perhaps, be indicated here. M. R. Schneck contributed most of the material to Part One; H. E. Garrett contributed the material to Part Two, and is responsible for the general plan of the book.

Acknowledgments to authors and publishers from whose work we have taken material are made at appropriate places in the text. We are indebted to the following colleagues and friends who read one or more chapters in manuscript: Professors R. S. Woodworth and C. J. Warden, and Dr. Otto Klineberg, of Columbia University; Professor J. W. Dunlap, of Fordham University; and Professors John

F. Walker and J. E. Caster, of the University of Arizona. Misses Franklyn Royer and Catherine Margan, of the University of Arizona, assisted in gathering material for Part One. Dr. Anne Anastasi, of Barnard College, and Mr. Benjamin H. Brown, of Columbia, rendered well-nigh indispensable service in many ways. We are grateful to Professor Gardner Murphy, the editor of the series, for many helpful suggestions.

H. E. GARRETT  
M. R. SCHNECK

# TABLE OF CONTENTS

## PART ONE

### *The Measurement of Simpler Functions*

I. TESTS OF PHYSICAL AND SENSORY CAPACITY	3
II. TESTS OF MOTOR ABILITY AND MECHANICAL APTITUDE	38
III. TESTS OF PERCEPTION AND ATTENTION	65
IV. TESTS OF LEARNING, ASSOCIATION AND MEMORY	92

## PART TWO

### *The Measurement of Complex Functions*

I. VERBAL OR LINGUISTIC TESTS OF "GENERAL INTELLIGENCE"	3
II. PERFORMANCE AND NON-LANGUAGE TESTS OF GENERAL MENTAL ABILITY	69
III. THE MEASUREMENT OF PERSONALITY AND TEMPERAMENT	102
IV. TESTS IN SPECIAL FIELDS	167
V. SOME APPLICATIONS OF PSYCHOLOGICAL TESTS	185
INDEX OF SUBJECTS	225
INDEX OF NAMES	231



## LIST OF FIGURES

## PART ONE

1. STADIOMETER, OR HEIGHT STAND	8
2. WET SPIROMETER	12
3. HEAD CALIPERS	15
4. HAND DYNAMOMETER	17
5. MCCALLIE VISION TESTS (ILLITERATES)	27
6. SEASHORE AUDIOMETER	32
7. TAPPING TEST	39
8. STEADINESS TEST	44
9. COÖRDINATION TEST (THREE HOLE)	47
10. AIMING OR TARGET TEST	49
11. TRACING BOARD	51
12. STENQUIST ASSEMBLING TESTS, SERIES 1	56
13. WHIPPLE DISC TACHISTOSCOPE	67
14. LETTER CANCELLATION TEST	72
15. CARD SORTING BOX	77
16. O'CONNOR WIGGLY BLOCK TEST	83
17. MINNESOTA PAPER FORM BOARD TEST	85
18. WITMER CYLINDER TEST	87
19. WOODWORTH-WELLS DIGIT-SYMBOL SUBSTITUTION TEST	94
20. PETERSON MENTAL MAZE	96
21. STYLUS MAZE (KLINE)	101
22. MIRROR DRAWING TEST (WHIPPLE)	102
23. FREE ASSOCIATION TEST (KENT-ROSANOFF)	109
24. LOGICAL MEMORY TEST AT YEAR 10 (STANFORD-BINET)	129
25. THE MARBLE STATUE—WHIPPLE [77]	129
26. PSYCHOGRAPH	136

## PART TWO

1. HYPOTHETICAL GROWTH CURVES WHICH GIVE A CONSTANT I.Q.	15
2. HYPOTHETICAL GROWTH CURVES WHICH GIVE A CONSTANT I.Q.	15



3. MENTAL GROWTH ON THE BINET TESTS MEASURED IN EQUAL UNITS (THURSTONE [69])	20
4. MENTAL GROWTH ON THORNDIKE'S CAVD INTELLIGENCE EXAMINATION MEASURED IN EQUAL UNITS (THORNDIKE [68])	21
5. DISTRIBUTION OF I.Q.'s (STANFORD-BINET) IN THE GEN- ERAL POPULATION	26
6. DISTRIBUTION OF GENERAL INTELLIGENCE (ARMY ALPHA SCORES) FOR WHITE, NATIVE-BORN, ENLISTED MEN	34
7. AGE-PROGRESS CURVES FOR 37,069 CHILDREN UPON THE NATIONAL INTELLIGENCE TEST	48
8. AGE-PROGRESS CURVES FOR 25,226 CHILDREN UPON THE OTIS GROUP INTELLIGENCE SCALE, ADVANCED EXAMINA- TION	49
9. PINTNER-PATERSON PERFORMANCE TESTS	77
10. PORTEUS MAZE TESTS FOR AGES 8, 9, 10, 11, 12, AND 14	84
11. FERGUSON FORM BOARDS	85
12. A SPECIMEN PAGE FROM THE ARMY BETA GROUP EXAMINA- TION	93
13. PINTNER-CUNNINGHAM PRIMARY MENTAL TEST	95
14. PERSONALITY RATING SCALE (AMERICAN COUNCIL ON EDU- CATION [7])	107
15. OVERLAPPING OF FREQUENCY DISTRIBUTIONS	195

■

# PART ONE

■



## CHAPTER I

### TESTS OF PHYSICAL AND SENSORY CAPACITY

#### I. PHYSICAL TESTS

WHILE the chief interest of the experimental psychologist in physical tests has been as a means of studying the relation of mental and physical characteristics, such tests have wide and diverse applications in other fields. To the anthropologist and sociologist, physical tests have long been useful in studying individual and group differences, as well as the effects of environmental factors upon bodily growth and general health. Physical tests are valuable to the applied psychologist engaged in vocational and educational guidance; and especially useful, also, to the clinical psychologist in dealing with delinquent, incorrigible and mentally backward individuals.

Many recent investigations have shown the extent to which physical deficiencies may affect social and emotional adjustments, as well as achievement in school. This knowledge has greatly stimulated the use of physical tests in schools and colleges. Interest in physical measurement has been aroused, too, by a movement among certain German psychologists and psychiatrists in which body measurement plays an important part. This movement has led to a "psychology of types," the purpose of which is to establish significant relations between body type and mental, emotional and temperamental make-up. Endocrinology has also contributed new impetus to the study of physical growth and physiological conditions. The rôle of the endocrine glands in personality growth and personality changes has opened up a new and fascinating field in which physical measures are extremely important.

The tests described in this section represent a selection from many existent tests of physical and physiological efficiency. Tests of narrow functions, specialized tests, and those of little value to the psychologist, have been omitted. Measurements of brightness discrimination,

NOTE: References in italics are to Part Two.

of skin and pain sensitivity, of weight discrimination and the like are not included. A comprehensive discussion of many of these tests will be found in Whipple (53).<sup>1</sup>

The following tests will be considered in some detail:

- Height
- Weight
- Lung capacity and vital index
- Cephalic index
- Strength of grip
- Strength of back
- Strength of legs
- Strength of arms
- Pulse rate

Methods of giving these tests, and some of the important results attained by their use, will be presented. It will also be shown how a group of physical measures may be combined to give an index of general physical efficiency. This index was devised by Rogers (39) as a measure of athletic ability. It is, however, a good indication of general bodily fitness, and hence is useful in other connections.

### 1. Height and Weight

Height is one of the most commonly measured physical characteristics, and correlations between it and other physical and mental traits have been sought in many studies (46). Besides its use as a separate measure, height enters into several indices which seek to define physical types. Naccarati (30), for instance, made use of height in his effort to discover a relationship between general intelligence and physical constitution. As a measure of constitution, Naccarati employed a *morphologic index*, which was obtained by adding together the length of one arm and one leg, and dividing this sum by the volume of the trunk. The ratio of height to weight was found to have a correlation of .75 with the morphologic index in one group of fifty college students; and .70 in a second group of seventy-five students. On the basis of these correlations, Naccarati took the height-weight ratio to be a fairly satisfactory substitute for the morphologic index. Hence, in many of his groups, he correlated measures of general intelligence with the height-weight ratio rather

<sup>1</sup>Numbers in parentheses throughout refer to the bibliography at the end of each chapter.

than with the less easily obtained morphologic index. Naccarati's major findings were as follows:

(a) The correlation between intelligence and the morphologic index was .75 in a group of seventy-five college students. The measure of intelligence was the Thorndike Entrance Examination, the reliability of which, according to Peatman (34), is .83.

(b) The correlation between the height-weight ratio and the Thorndike Examination was .23 in a group of 221 college students, including the seventy-five mentioned in (a).

(c) In the same group of 221 subjects, the correlation of height alone with the Thorndike Examination was .04.

These results point to a closer relationship between intelligence tests and the morphologic index than between intelligence and either height or the height-weight ratio. But in no case is the relationship high. Sheldon (42) employed the morphologic index in an investigation at the University of Chicago, and obtained much the same results as Naccarati. Sheldon used as his measure of intelligence the Psychological Examination of the American Counsel on Education (1924 Edition) constructed by L. L. Thurstone. His group consisted of 450 college students. A correlation of .14 was obtained by Sheldon between the morphologic index and general intelligence test scores, which is even lower than the result obtained by Naccarati. It is possible, and even probable, that in a group less highly selected than college students, the correlation between body type and measures of general ability would be higher than the figures cited. It is worth noting, however, that Sheldon found a correlation between estimates of intelligence (based upon ten minutes' acquaintance) and grades in the University of Chicago of .21; while the correlation between morphologic index and grades was .11.

Sheldon (43) has reported another interesting study in which the morphologic index was used. Taking his cue from the fact that Kretschmer (23) found a relationship between certain types of physique and mental traits, Sheldon looked for a relationship between the morphologic index and social traits. This possibility is supported, superficially, by the resemblance of Kretschmer's types to Naccarati's types. Sheldon obtained ratings of 155 college freshmen, in the following traits: sociability, perseverance, leadership, aggressiveness and emotional excitability. Re-ratings of twenty-eight subjects a month later showed a rating reliability ranging from .82

to .93. Of sixty correlations calculated between social ratings and either the morphologic index or its components, twelve are above .10, two are above .20, and seventeen are negative. The highest positive correlation was .24, and the highest negative correlation —.22. These correlations are low, but Sheldon suggests that they are too high and too numerous to be due entirely to chance. It is doubtful, however, whether the relationship of the morphologic index to social traits is closer than is its relationship to measures of intelligence or to scholastic aptitude.

Kretschmer (23) believes that he has discovered a decided relationship between physical type and temperamental make-up. He has described four physical types, as follows: (a) the asthenic physique, which is tall, slender, and relatively feeble; (b) the athletic, strong, well-muscled, vigorous; (c) the pyknic, inclined to obesity and usually short in stature; (d) the dysplastic, an intermediate group whose abnormalities usually arise from endocrine disturbance. Kretschmer's temperamental types are (a) the cycloid, which is akin to the manic or excited phase of manic-depressive psychosis and (b) the schizoid, which is characterized by unsociable, shy and introverted behavior, and is akin to dementia præcox. Kretschmer's data showing the relation between the physical and temperamental types are impressive when taken at face value:

RELATION BETWEEN PHYSICAL TYPES AND MENTAL DISEASE TYPES

Physical Type	No. of Cycloids	No. of Schizoids
Asthenic . . . . .	4	81
Athletic . . . . .	3	31
Pyknic. . . . .	58	2

The accuracy of this table depends, of course, upon the validity of Kretschmer's classification of the mental disease types. It is possible that this classification was influenced by Kretschmer's psychiatric theories. In any event, the scheme is based upon extreme, *i.e.*, pathological, cases; and it is important to know to what extent the classification applies to non-pathological cases.

Most of the correlations between height and intelligence are negligible and practically all of them are low. Naccarati's low correlation has already been quoted (p. 5). Sommerville (46), in a group of ninety-eight college students, correlated height with the Thorndike Entrance Examination. When variability arising from age was held constant, the coefficient was .16. Abernethy (1) has studied the relation between height and mental age (Stanford-Binet), her sub-

jects being 120 girls between six and twelve years of age. When the age factor was held constant statistically, the correlation between height and mental age was .34, which is fairly high. In an older group, ages thirteen to seventeen, the correlation, when calculated for each age group separately, ranged from .01 to .25, and the probable errors from .08 to .12. Gittings (15) has studied the relation of mental to physical traits in a group of seventy-five freshmen women in the University of Arizona. She found the correlation between height and scores in the Army Alpha Test to be .18, which is, of course, statistically unreliable. In general, these results indicate a negligible correlation between height and intelligence.

Correlations between weight and measures of intelligence are no more impressive than are those for height and intelligence. Naccarati (30), in a group of 221 college students, reported the correlation between weight and the Thorndike Entrance Examination to be —.18. Naccarati and Guinzberg (31), using the same intelligence examination with 252 male college students below the age of twenty-one, found the correlation with weight to be —.02. Murdoch and Sullivan (29) have studied the relationship between intelligence measures and weight in a group of 600 children, six to eighteen years old. Standard tests were used, *viz.*, the Otis Primary Intelligence Test, the National Intelligence Test and the Terman Group Intelligence Test. For the group as a whole, the correlation between intelligence, as expressed in terms of the intelligence quotient, and weight was .16. Subdividing the group into three divisions, Murdoch and Sullivan report the following correlations: Pre-adolescents, .17 ( $N = 171$ ); adolescents, .16 ( $N = 262$ ); post-adolescents, .14 ( $N = 171$ ). This study included boys and girls, but the correlations were not appreciably altered when coefficients were computed separately for each sex. Gates (14) has substantiated these results with young children.

The lack of correlation between height and weight and measures of general mental ability does not mean necessarily that these physical measures have little value for psychology. After all, height and weight are rather gross measures of physical status, subject to many influences, and one could hardly expect them to show a high or consistent relationship to measures of mental ability. The most valuable information for psychologists, found by those investigators who have studied height and weight in various connections, has been summarized by Whipple (53) as follows:



(a) Boys continue to increase in both height and weight later than do girls. Girls grow most rapidly between ten and fifteen years; boys between twelve and seventeen years.

(b) During the pre-pubertal period (eleven to thirteen years) girls are taller and heavier than boys of the same age.

(c) During the period of adolescence, children do not all grow at the same rate. Variations in weight are largest during this period.

(d) First-born children tend to be taller and heavier than later-born children, but the difference is probably not significant.

(e) Boys in the public schools are in general taller and heavier than boys in truant schools.

(f) In an extensive survey, Goddard (16) found that feeble-minded children are shorter and lighter than normal children of the same age. The greater the mental defect, the greater the divergence from normal height and weight.

(g) The correlation between height and weight will vary from .50 to .75, depending upon the size and the age level of the group.

Directions for the accurate measurement of height and of weight follow. Average measurements, for boys, girls, men and women are given in Tables I to IV.

#### Test 1. Height

*Apparatus:* Stadiometer, obtainable from the C. H. Stoelting Company, Chicago, Illinois, or from the Marietta Apparatus Company, Marietta, Ohio.

*Method:* The subject (S), with shoes removed, stands on the box of the stadiometer, with heels, shoulders and back of the head in light contact with the upright rod. The chin should not be raised or lowered. The experimenter (E) brings the sliding horizontal bar down until it rests squarely, without pressure, on the head. The reading is taken directly from the scale on the upright rod.

*Record:* Measurement may be taken in centimeters or in inches.

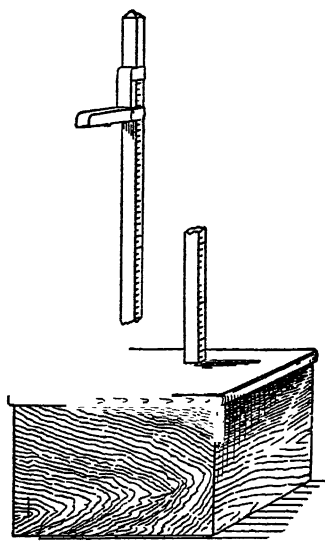


Figure 1.—STADIOMETER, OR  
HEIGHT STAND  
(Reproduced by courtesy of the  
C. H. Stoelting Co.)

*Alternative Method:* If the stadiometer is not available, a scale may be marked off on a wall or some other upright surface. The surface on which S stands should be perfectly level.

## Test 2. Weight

*Apparatus:* Accurate scales, preferably graduated both in pounds and kilograms. Obtainable from the C. H. Stoelting Co., Chicago, Illinois, or the Marietta Apparatus Co., Marietta, Ohio.

*Method:* For exact results, weight should be taken without clothing. But

### HEIGHT AND WEIGHT TABLES

TABLE I

#### WEIGHT-HEIGHT-AGE TABLE FOR BOYS

! By B. T. Baldwin and T. D. Wood

(From H. L. Smith and W. W. Wright [45])

Height in Inches	Average Weight in Pounds	Age																
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19		
38	34	34	34															
39	35	35	35															
40	36	36	36															
41	38	38	38	38														
42	39	39	39	39	39													
43	41	41	41	41	41													
44	44	44	44	44	44													
45	46	46	46	46	46	46												
46	48	47	48	48	48	48												
47	50	49	50	50	50	50	50											
48	53		52	53	53	53	53											
49	55		55	55	55	55	55	55										
50	58		57	58	58	58	58	58	58									
51	61			61	61	61	61	61	61	61								
52	64			63	64	64	64	64	64	64	64							
53	68			66	67	67	67	67	67	68	68							
54	71				70	70	70	70	71	71	72							
55	74				72	72	73	73	74	74	74							
56	78				75	76	77	77	77	78	78	80						
57	82					79	80	81	81	82	83	83						
58	85					83	84	84	85	85	86	87						
59	89						87	88	89	89	90	90	90					
60	94						91	92	92	93	94	95	96					
61	99							95	96	97	99	100	103	106				
62	104							100	101	102	103	104	107	111	116			
63	111							105	106	107	108	110	113	118	123	127		
64	117								109	111	113	115	117	121	126	130		
65	123								114	117	118	120	122	127	131	134		
66	129									119	122	125	128	132	136	139		
67	133									124	128	130	134	136	139	142		
68	139										134	134	137	141	143	147		
69	144										137	139	143	146	149	152		
70	147										143	144	145	148	151	155		
71	152										148	150	151	152	154	159		
72	157											153	155	156	158	163		
73	163												157	160	162	164	167	
74	169													160	164	168	170	171

most tables of norms, including that on p. 9, are based upon weight taken in ordinary indoor clothing and shoes. Fairly accurate figures for weight without clothes may be obtained by subtracting 5 per cent. of the gross weight for children, 6 to 8 per cent. for men, and 4 to 5 per cent. for women, depending upon the season.

*Note:* The following remarks are offered by Wood (57): "Up to twenty years of age, it is advantageous for health to weigh as much as the standard in the table for height and age. Above the age of thirty, *over-weight* is decidedly disadvantageous to health. For a person above thirty, the best weight standard is given by the table in the thirty-to-thirty-four-year column."

TABLE II  
WEIGHT-HEIGHT-AGE TABLE FOR GIRLS  
By B. T. Baldwin and T. D. Wood  
(From H. L. Smith and W. W. Wright [45])

Height in Inches	Average Weight in Pounds	Age																		
		5	6	7	8	9	10	11	12	13	14	15	16	17	18					
38	33	33	33																	
39	34	34	34																	
40	36	36	36	36																
41	37	37	37	37																
42	39	39	39	39																
43	41	41	41	41	41															
44	42	42	42	42	42															
45	45	45	45	45	45	45														
46	47	47	47	47	48	48														
47	50	49	50	50	50	50	50													
48	52		52	52	52	52	53	53												
49	55		54	54	55	55	56	56												
50	58		56	56	57	58	59	61	62											
51	61			59	60	61	61	63	65											
52	64			63	64	64	64	65	67											
53	68			66	67	67	68	68	69	71										
54	71				69	70	70	71	71	73										
55	75					74	74	74	75	77	78									
56	79					76	78	78	79	81	83									
57	84						82	82	82	84	88	92								
58	89						84	86	86	88	93	96	101							
59	95						87	90	90	92	96	100	103	104						
60	101						91	95	95	97	101	105	108	109	111					
61	108							99	100	101	105	108	112	113	116					
62	114							104	105	106	109	113	115	117	118					
63	118								110	110	112	116	117	119	120					
64	121									114	115	117	119	120	122	123				
65	125										118	120	121	122	123	125	126			
66	129											124	124	125	128	129	130			
67	133												128	130	131	133	135			
68	138												131	133	135	136	138	138		
69	142													135	137	138	140	142		
70	144														136	138	140	142	144	
71	145															138	140	142	144	145

TABLE III  
WEIGHT-HEIGHT-AGE TABLE FOR WOMEN  
(From T. D. Wood [57])

Height in Inches	Age									
	19	20	21-22	23-24	25-29	30-34	35-39	40-44	45-49	50-54
58	104	106	108	110	113	116	119	123	126	129
59	106	107	109	112	115	118	121	125	128	131
60	112	112	113	115	117	120	123	127	130	133
61	116	116	116	118	119	122	125	129	132	135
62	118	118	119	120	121	124	127	132	135	138
63	120	121	122	123	124	127	130	135	138	141
64	123	124	125	126	128	131	134	138	141	144
65	126	127	128	129	131	134	138	142	145	148
66	130	131	132	133	135	138	142	146	149	152
67	135	135	135	137	139	142	146	150	153	156
68	138	138	139	141	143	146	150	154	157	161
69	142	142	142	145	147	150	154	158	161	165
70	144	144	145	148	151	154	157	161	164	169
71	146	147	149	151	155	157	160	164	168	173
72	150	152	154	156	158	161	163	167	171	176

TABLE IV  
WEIGHT-HEIGHT-AGE TABLE FOR MEN  
(From T. D. Wood [57])

Height in Inches	Age										
	19	20	21-22	23-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59
60	111	112	114	118	122	126	128	131	133	134	135
61	116	117	118	121	124	128	130	133	135	136	137
62	122	123	124	125	126	130	132	135	137	138	139
63	127	128	128	129	131	133	135	138	140	141	142
64	130	131	132	134	135	136	138	141	143	144	145
65	134	135	136	137	138	140	142	145	147	148	149
66	139	140	141	142	143	144	146	149	151	152	153
67	142	143	144	145	146	148	150	153	155	156	158
68	147	148	149	150	151	152	153	158	160	161	163
69	152	153	154	155	156	158	160	163	165	166	168
70	155	156	157	158	159	162	165	168	170	171	173
71	159	160	161	162	164	166	170	174	176	177	178
72	163	164	165	166	168	172	176	180	182	183	184
73	167	168	169	171	173	178	183	186	188	190	191
74	171	172	174	176	179	184	189	193	195	197	198
75	175	175	178	181	184	190	195	200	202	204	205
76	178	180	183	186	189	196	201	206	209	211	212
77	183	185	188	191	194	201	207	212	215	217	219

## 2. Vital Capacity and Vital Index

Vital capacity is defined as the greatest quantity of air that can be exhaled from the lungs after a maximal inhalation. Together with weight, it has long been considered a valuable index of physical condition. The vital index is obtained by dividing vital capacity, in

cubic centimeters, by weight, in kilograms. Weight gives us an indication of bodily size, and therefore of the quantity of tissue to be supplied with blood by the circulatory system. Since the major function of breathing is that of oxidizing the blood, the ratio of vital capacity to weight is presumably a measure of the efficiency of the oxidation process. Some writers prefer to calculate the vital index by substituting height for weight. Among these are Collins and Howe (6), and A. H. Turner (52), the latter finding that there is a

closer relationship between vital capacity and height than between vital capacity and weight. But the usual calculation employs weight. Rogers (39) reports that though the reliability of lung capacity is high, it ranks lowest of all physical capacity tests as a measure of general athletic ability. Lung capacity is, however, used by Rogers in the calculation of his strength index.

As in the case of height and of weight, such correlations as exist between vital capacity or the vital index and intelligence test scores are low. DeBusk (7) found that, for 105 school children, differences between normal and pedagogically retarded children are greater for vital capacity than for height and weight. Whipple (53) quotes Goddard as finding that the lung capacity of feeble-minded children is below that of normal children. But Naccarati (30), in a group of 136 college students,

Figure 2.—WET SPIROMETER  
(Reproduced by courtesy of  
the C. H. Stoelting Co.)

found the correlation between lung capacity and either the Otis Intelligence Test or the Thorndike Entrance Examination to be—.11. Sommerville (46), correlating the Thorndike Examination with measures of lung capacity for ninety-eight college students, arrived at a coefficient of .16, while the correlation of the same examination with vital index for ninety-six college students was .11. Gittings (15), for seventy-five freshmen women in Arizona, obtained a correlation of .18 between vital capacity and Army Alpha scores.

Vital capacity and the vital index are useful in the calculation of

morphologic indices, in physical education, and in medicine. They enter into an index of build or stature proposed by Dreyer and Hanson (9), into the respiratory-height coefficient of Williams (54), and into Rogers' Physical Fitness Index (39). Turner (52) found, after studying the records of Wellesley freshmen, that students of high vital capacity not only have better medical records than do those of low vital capacity, but that they are also better nourished and superior in college athletics.

### Test 3. Vital Capacity

*Apparatus:* Wet spirometer and a supply of sterilized detachable wood or glass mouthpieces. If a wet spirometer is not available, a dry spirometer may be used. The dry spirometer is less accurate than the wet; and readings from the two are not strictly comparable. Spirometers may be purchased from the Narragansett Machine Co., Providence, Rhode Island; from the C. H. Stoelting Co., Chicago, Illinois, and from the Marietta Apparatus Co., Marietta, Ohio.

*Method:* Use a fresh mouthpiece for each subject. Have S stand erect, take as deep a breath as he can, and exhale fully into the spirometer.

TABLE V  
NORMS OF VITAL CAPACITY, IN CUBIC CENTIMETERS  
(Whipple, after Smedley [53])

Age	Boys	Girls	Age	Boys	Girls
6	1,023	950	13	2,108	1,827
7	1,168	1,061	14	2,395	2,014
8	1,316	1,165	15	2,697	2,168
9	1,469	1,286	16	3,120	2,266
10	1,603	1,409	17	3,483	2,319
11	1,732	1,526	18	3,655	2,343
12	1,883	1,664			

This should be done slowly and care should be taken that none of the air escapes. Three trials should be taken.

*Record:* Readings are taken from the scale, in cubic inches and in liters. Convert liters into cubic centimeters by multiplying by 1,000 (1,000 cubic centimeters = 1 liter). Readings on the dry spirometer are given only in cubic inches.

*Norms:* Vital capacity varies with age, sex, height, weight, chest circumference and habits of life. It is difficult to give definite norms because of this complexity. Smedley's norms, taken from Whipple (53), are given in Table V. Wilson and Edwards (55), who measured the vital capacity of 362 children, six to sixteen years of age, offer the following standard of vital capacity; namely, that vital capacity should equal 15.5 cubic centimeters for each centimeter in height.

### 3. Cephalic Index

This index is computed from measurements of the length and width of the head in millimeters, as follows:

$$\text{Cephalic Index} = \frac{\text{Width} \times 100}{\text{Length}}$$

A customary, but not invariable, classification of head shapes is the following:

	Index
Long-headed or dolichocephalic . . . . .	Below 75
Medium head or mesocephalic . . . . .	75 to 80
Broad head or brachycephalic . . . . .	Above 80

Special interest attaches to this index, since it has been popularly supposed that there must be a direct relationship between size and form of head and intelligence. This relationship, if established, would have bearing upon racial and national differences in head form. Kroeber (24) states that there are no typical racial head forms, but that sub-types can be distinguished. Dixon (8) holds that the form of the head is the most permanent and distinctive of racial traits. Boas (2) presents evidence indicating that the form of the head is modifiable, environment causing a drop in the cephalic index of children of broad-headed Hebrew immigrants in New York, and a rise in the cephalic index of children of long-headed Sicilian immigrants. Huntington (19) and Radosavljevich (37) take issue with Boas, insisting that the latter's investigations were faulty in method. The question is not settled, and interest in cephalic measurements is still active.

The relation between head form and intelligence has interested many psychologists. In extreme cases, very small heads (microcephaly), and very large heads (hydrocephaly) are associated with low grades of intelligence (51); but within normal limits the correlation between cranial measurements and intelligence is not high. A full discussion of this relationship is given by Paterson (32), whose summary of the experimental work done along these lines is convincing. A much-quoted study is that of Pearson (33), who correlated head measurements with teachers' estimates of intelligence and with scholastic standing for 1,011 Cambridge students and for about 2,300 twelve-year-old boys and 2,200 twelve-year-old girls.

The correlations of the cephalic index with the estimates of intelligence in these groups were  $-.06$ ,  $-.04$ , and  $.07$ , respectively. The correlations between length of head and width of head and the same measures of intelligence are all positive, but do not exceed  $.14$ . Sommerville (46), for 105 college students, reports a correlation of  $-.01$  between cephalic index and the Thorndike Entrance Examination. In the same report Sommerville found a correlation of  $.10$  between cranial capacity and the Thorndike Examination. His correlations for intelligence and head length and head width are in close agreement with those of Pearson. Murdoch and Sullivan, in the study quoted on p. 7, found a correlation of  $.22$  between head diameter and I.Q. Another study, by Reid and Mulligan (38), reports a correlation of  $.08$  between cranial capacity and records of scholastic standing for 449 students (male) at Aberdeen University.

Many other studies have been made in this field, the usual finding being a slight positive correlation between head measurements and intelligence-test results. Hull (18) has suggested that possibly a combination of cranial and physiognomic measurements might be worked out which would yield a substantial correlation with measures of intelligence or with scholastic aptitude. Quoting a study by Sherman (44) done under his direction, Hull states that an index derived from two measurements of facial angles and three cranial measurements, made upon seventy-eight freshmen in an engineering college, gave a correlation of  $.50$  with academic marks in scientific and engineering subjects. Paterson points out (32) that this result is far out of line with the finding of Pearson and other investigators. Nevertheless, Hull's and Sherman's study is valuable, since it indicates the distinct possibility of combining various head dimensions in such a way as to get a significant correlation with measures of intellectual activity.

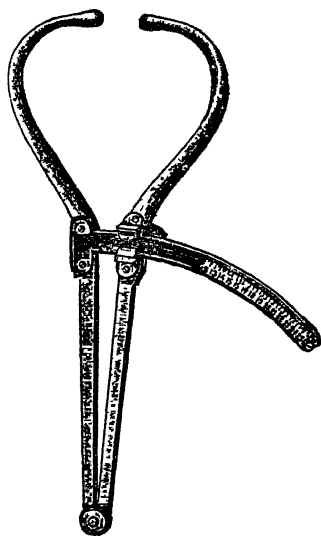


Figure 3.—HEAD CALIPERS  
(Reproduced by courtesy of the  
C. H. Stoelting Co.)



#### Test 4. Head Measurements

*Apparatus:* Head calipers, obtainable from the C. H. Stoelting Co., Chicago, Illinois.

*Method:* To measure the length of the head, S is seated, and the left tip of the calipers placed against the glabella (the space between the eyebrows). The right tip is placed against the farthest projection at the back of the head. Hold the tips firmly between thumb and forefinger, letting the upper part of the calipers rest against the chest. When the greatest length seems to have been secured, note the reading. Remove the calipers. Tighten the thumbscrew, and apply the calipers again to make sure that the reading is correct.

To obtain the width of head, stand behind S and apply one tip of the calipers to the left side of the head and the other to the right side, just above the ears. Move the calipers up vertically until the maximal reading is obtained. Tighten the thumbscrew at this point and apply the calipers again for verification.

*Record:* Record the readings in millimeters. Compute the cephalic index from the formula on p. 14.

*Norms:* Cranial measurements vary with age, race and sex. Wissler (56) sums up the work of previous investigators as follows: during school life the heads of boys and girls become longer with age; the fall in the index during growth is slight for North European children and relatively great for the round-headed Chinese and Japanese children. Adult women tend to be rounder-headed than men. Wissler's own results taken from about 9,000 children in Hawaii indicate that in early life boys have rounder heads than girls, but that later girls become rounder-headed than boys. Whether these sex and race differences are correlated with mental traits is problematical. Stockard (49) has discussed cephalic types with reference to their characteristics, their geographical distribution and age modifications.

#### 4. Strength Tests

Many investigators recognize the fact that dynamic tests are, in general, superior to static indices of physical efficiency, such as height and weight (3). For certain kinds of athletic and gymnastic work it is often desirable to know the amount of muscular strength possessed; and this is also true in many types of gainful labor. Muscular strength is an important factor in estimates of general physical condition (39), and many studies of the relation between muscular strength and intelligence have been made. A few typical investigations are cited below.

Whipple (53) cites Smedley, Schuyten and Carman as having found evidence of positive correlation between strength of grip and intelligence. Woolley and Fischer (58) gave a variety of mental and

motor tests to about 700 boys and girls, fourteen and fifteen years old, and found practically no correlation between hand grip and the mental tests. Rudisill (40), in a group of forty college students (thirteen men, twenty-seven women) reported a correlation of .39 between grip and the Army Alpha Test for the men, and  $-.21$  for the women. These are fairly substantial correlations, but the coefficients are not very reliable in view of the small number of subjects. Garfiel (13), working with thirty-two sophomore girls in Barnard College, found that the correlation of the Army Alpha Test with strength of back was .23; with hand grip, .06; with the leg dynamometer,  $-.01$ . The probable errors were .10, which makes all of these coefficients unreliable. Johnson (20) investigated the relationship between strength of grip and mental age (Stanford-Binet Scale), her subjects being 262 boys and girls, three to thirteen years old. The correlation was .71, which is very high. But Johnson shows that height and weight markedly influence the strength score, and that the correlation between grip and chronological age for ninety-eight cases was .77. When the influences of height, weight and chronological age are held constant statistically, Johnson's resulting correlation between mental age and strength of grip has the negligible value of .03.

Monahan and Hollingworth (28) report an interesting study made upon two groups of children in a New York public school. The experimental group consisted of forty-five boys and girls with a mean I.Q. of 152 and a mean chronological age of 135.2 months. The control group contained forty-five boys and girls, matched with the experimental group for sex, age and racial stock, but with a mean I.Q. of about 100. In the experimental group the mean score for hand grip was 25 kilograms, with a mean deviation of 3.29; while in the control group the score was 23.4 kilograms, with a mean deviation of 3.56. Monahan and Hollingworth conclude that the difference between the groups is real, since their figures show that the absolute difference between the groups, although small (1.6 kilograms), is 3.33

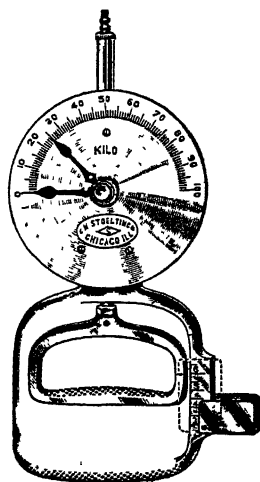


Figure 4.—HAND  
DYNAMOMETER  
(Reproduced by courtesy  
of the C. H. Staeling  
Co.)

times the P.E. of the difference. Superior children, therefore, appear on the whole to be slightly stronger physically than average children. There is little justification, however, for concluding from the results of this study that muscular strength bears any significant relation to measures of mental ability. An excellent summary of the work done upon this and other relations between mental and physical traits is given by Stalnaker (47).

#### Test 5. Strength of Grip

*Apparatus:* Hand dynamometer, preferably the Smedley type, obtainable from the C. H. Stoelting Co., Chicago, Illinois.

*Method:* Have S grip the instrument so that the second phalanges of the fingers press against the inner handle. Close the clutch in order to hold the inner handle firm. Have S grip the two handles firmly, raise the dynamometer until level with his head, and then bring it down quickly, at the same time exerting maximal pressure on the handles. Take three trials for each hand, alternating from right hand to left.

*Record:* The score is the best of three trials for each hand. The scale is graduated in terms of kilograms.

TABLE VI  
NORMS FOR STRENGTH OF GRIP (IN KILOGRAMS)  
(From Whipple, after Smedley [53])

Age	Boys		Girls	
	Right Hand	Left Hand	Right Hand	Left Hand
6 . . . . .	9 21	8 18	8 36	7 74
7 . . . . .	10 74	10 11	9 88	9 24
8 . . . . .	12 41	11 67	11 16	10 18
9 . . . . .	11 34	13 47	12 77	11 97
10 . . . . .	16 52	15 59	14 65	13 72
11 . . . . .	18 85	17 72	16 54	15 52
12 . . . . .	21 24	19 71	18 92	17 78
13 . . . . .	24 14	22 51	21 84	20 39
14 . . . . .	28 42	26 22	24 79	22 92
15 . . . . .	33 39	30 88	27 00	24 92
16 . . . . .	39 37	36 39	28 70	26 56
17 . . . . .	44 74	40 96	29 56	27 13
18 . . . . .	49 28	45 01	29 75	27 66

#### Test 6. Strength of Back

*Apparatus:* Back and leg dynamometer, obtainable from the C. H. Stoelting Co., Chicago, Illinois.

*Method:* Have S stand upon the foot-rest of the dynamometer. Adjust the length of the chain so that S's body is bent forward at a comfortable angle, usually about 60°, knees unbent. Then have S take a deep breath and lift with a maximal effort, using the back and arms without bending the knees.

*Record:* The score, in pounds or in kilograms, is the better record of two trials. An interval for rest is allowed between trials.

#### Test 7. Strength of Leg

*Apparatus.* The same as for the back.

*Method:* Have S stand upon the foot-rest of the dynamometer and grasp the handle, with knees bent, and body and head held erect. The handle should rest against the thighs. Instruct S to take a deep breath and exert a maximal lift, confining the effort to his legs.

*Record:* The same as for strength of back.

#### Test 8. Push-ups

*Apparatus:* Standard parallel bars, obtainable from any gymnasium supply house.

*Method:* The bars should be adjusted at about the height of the subject's shoulders, or a little lower. Have S stand between the bars, grasp one with each hand, and lift his body until his arms are straight. Instruct him to lower his body by bending his arms; then raise himself by straightening his arms, and continue in this way to lower and raise himself until he can do so no longer.

*Record:* The score is the number of complete push-ups. After a rest, the test may be repeated, and the best score taken as the final record.

#### Test 9. Pull-ups

*Apparatus:* Standard gymnasium eight-inch rings, obtainable from any gymnasium supply house.

*Method:* The rings are suspended from the ceiling or from a horizontal ladder. The test is the standard performance of "chinning."

*Record:* The score is the number of complete pull-ups. S pulls himself up and lowers himself alternately until he can do so no longer. A second trial may be taken after a rest period.

### 5. A General Index of Physical Capacity

One of the more ambitious efforts to develop an index of physical efficiency is that of Rogers (39), who was looking particularly for a workable basis upon which to group contestants in athletic competition. He derived a Strength Index and an Athletic Index, based upon the 100-yard dash, the broad jump, the high jump, the shot-put, and skill in baseball, football, and basketball throws. The correlation between this Athletic Index and the Strength Index was .81, and the reliability of the Strength Index was .94, in a group of 136 boys in grades 7 to 12, inclusive. The Strength Index is computed as follows:

1. Lung capacity. The best record from two tests.
2. Right-hand grip. The best record from two tests, alternating with left-hand tests.

3. Left-hand grip. The best record from two tests, alternating with right-hand tests.
4. Score for strength of back.
5. Score for strength of legs.
6. Push-ups score. This is multiplied by one-tenth of the subject's weight plus height minus 60. Stating it as a formula:

$$\text{Number of Push-ups} \times \left[ \frac{\text{Weight}}{10} + (\text{Height} - 60) \right]$$

7. Pull-ups score. Apply to this score the formula used for push-ups.
8. The normal strength index is the total of these seven scores.

An illustrative case should make the computation clearer:

Assume that the following data have been collected from tests given to a boy:

Weight . . . . .	142 pounds
Height . . . . .	67 inches
Pull-ups score . . . . .	9
Push-ups score . . . . .	8
Strength of legs . . . . .	540 pounds
Strength of back . . . . .	320 "
Left-hand grip . . . . .	95 "
Right-hand grip . . . . .	100 "
Lung capacity . . . . .	222 cubic inches

Applying the formula for pull-ups:

$$9 \times \left[ \frac{142}{10} + (67 - 60) \right] = 190.8$$

Applying the formula for push-ups:

$$8 \times \left[ \frac{142}{10} + (67 - 60) \right] = 169.6$$

---


$$\text{Total} = 360.4$$

Add the scores for the last five tests above:

	360 4
	540 0
	320 0
	95 0
	100 0
	222 0
	<hr/>
Strength Index . . . . .	1637 4

Since the strength index is in part conditioned by age and weight, it is necessary to adjust the normal index by allowing for these fac-

tors. These adjustments, taken from Rogers (39) are supplied in Table VII.

TABLE VII  
STRENGTH INDEX NORMS AND MULTIPLIERS FOR CERTAIN  
AGES AND WEIGHTS

Age	Weight	Normal Strength Index	Weight Deviation Multiplier
11-6	85	741	9 7
11-9	85	748	9 6
12-0	85	755	9 5
12-3	85	763	9 4
12-6	85	771	9 3
12-9	90	825	9 3
13-0	90	834	9 5
13-3	95	890	9 8
13-6	100	948	10 3
13-9	100	958	11 0
14-0	100	968	11 8
14-3	105	1037	12 9
14-6	105	1047	14 0
14-9	110	1127	15 0
15-0	110	1137	15 9
15-3	115	1231	16 7
15-6	120	1323	17 4
15-9	125	1420	18 0
16-0	125	1431	18 5
16-3	130	1534	18 9
16-6	130	1544	19 3
16-9	135	1651	19 6
17-0	135	1662	19 9
17-3	135	1673	20 1
17-6	140	1786	20 3
17-9	140	1798	20 5
18-0	140	1810	20 5
18-3	145	1924	20 6
18-6	145	1935	20 6
18-9	145	1946	20 6
19-0	145	1958	20 6
19-3	145	1970	20 6

To compute a normal score for any age and weight from this table: (a) Find the norm for the given age; (b) compute the difference between the weight given for this age and the actual weight of the subject; (c) multiply this difference (in pounds) by the multiplier for the subject's age. If the subject's weight is above that shown for the norm, add this result to the norm; if it is below, subtract it. The result is the Strength Index norm for the age and weight of the subject.

It should be remembered that deviations from norms established by investigators who have used the tests described are to be regarded

as genuinely significant only when such deviations are quite marked. A norm is simply an average, and it is possible for an individual to depart fairly widely from the norm without being in any sense atypical. Measures of single traits are less significant than are the relationships of these traits to one another. As Rogers remarks, a boy may vary in weight and still be in satisfactory physical condition. But to vary widely in weight, without varying in strength and endurance also, may be significant. In such cases indices which express the relationships of two or more traits belonging to the same individual are especially valuable.

## 6. Pulse-rate

Pulse-rate is a functional test which is coming into wider use constantly. There are several varieties of this test; in all of them it is essential to take not only the absolute pulse-rate but also the rate at which the pulse-beat returns to normal after exercise. The tests differ in methods of administration, in the exercise prescribed, and therefore in the norms established. From among those available we have chosen the Michigan Pulse-Rate Test (27) because it is easy to administer, is scored simply, and requires a mild form of exercise suitable for the laboratory. This test may be used in a group, in which case the subjects count the pulse themselves, both before and after the exercise. This method is not recommended for general use, however, since the accurate count of the pulse is not always easy, especially after exercise.

### Test 10. Pulse-rate

*Apparatus:* An accurate watch with a second hand, or a stop-watch.

*Method:* 1. Normal pulse-rate: While S is standing quietly, count his pulse for one minute. Repeat until the same rate is found two or three times in succession. If there is a marked deviation from the normal steady beat, postpone the test.

2. Instruct S to take three steps per second, in place, for fifteen seconds, or a total of forty-five steps. The foot should be raised halfway to the height of the opposite knee. Be sure that this movement is done properly. Note the exact time of starting and stopping the exercise.

3. Beginning exactly thirty seconds after the exercise has stopped, count the pulse for twenty seconds and write down the result. Later, this figure is to be multiplied by three to give the rate per minute. Exactly sixty seconds after the cessation of exercise, count the pulse again for twenty seconds and repeat at intervals of one minute until the normal pulse-rate has returned.

*Record:* The following scheme of classification is proposed by the State Council of Physical Education for Michigan:

Time to Recover Normal Rate	Grade	Fitness
30 seconds	A	Fine
60 "	B	Good
120 "	C	Fair
180 "	D	Poor
Pulse slower after exercise than before	E	Undetermined

In the last case the test should be repeated provided such repetition would not be injurious to S, since the initial count may have been incorrect.

## II. SENSORY TESTS

The efficiency of the sense-organs is highly important to normal existence. Some writers have held that there is a direct relationship between sensory discrimination and school achievement (36); and retardation in school has been frequently traced to sensory deficiencies (4, 12). Sterling and Bell (48) tested 1,860 children, ages eight to seventeen, inclusive, of both sexes, in the schools of Washington, D. C., and Hagerstown, Maryland, using an audiometer to measure acuity of hearing. A significant loss of hearing was taken to be a loss of nine or more units on the audiometer scale, the normal hearing being taken as the standard. The results are too numerous to be given in detail, but the following is a fair summary:

1. From 1 to 2.4 per cent. of the children showed a significant loss of hearing in the better ear. When both left and right ears were considered, the percentage of loss ranged from 3.3 to 8.2.

2. Significant loss was shown for 2.9 per cent. of children who were over age for their school grades, and for 1.2 per cent. of children at age for their grades.

3. When we divide the children into three groups—those doing excellent school work, those whose work is satisfactory and those whose work is unsatisfactory—the percentages of individuals with significant loss in hearing are 1.6, 1.7, 2.2. As the character of the work grows poorer, the significant loss in hearing tends to become greater.

4. The Washington group (585 in number) was also divided into three groups on the basis of I.Q. The percentages of significant loss in hearing were:

I.Q. above average	0.6 per cent. of the children
I.Q. at average	1.6 per cent. of the children
I.Q. below average	3.7 per cent. of the children



It is probable that the deficiency in hearing reduced the intelligence quotients in many of these cases, especially if an intelligence test, which involves oral questions, was used.

Kempf and Collins (21) conducted a survey of 5,000 school children in Illinois, the ages ranging from six to thirteen and over. Intelligence quotients were determined from the Otis Primary Intelligence Test for the first three grades, and from the Haggerty Intelligence Test for grades four to eight. Retarded children were tested individually, the Stanford Revision of the Binet Scale being used. No correlations are given but the results show some relationship between low I.Q. and defects in acuity of vision and of hearing.

Two methods of measuring sensory capacity are in general use. In the first method, we try to find the minimal stimulus which can be perceived by the subject. This minimum is called the absolute threshold or the limen of sensitivity. The faintest sound that the subject can hear, the lightest touch that he can feel, or the dimmest light that he can see is the limen of hearing, touch or vision, respectively. In the second method, we determine the smallest difference between two stimuli which can be distinguished by the subject. This minimal difference has been termed the just noticeable difference, the differential threshold, or the difference limen. To illustrate, we may determine the smallest difference that the subject can perceive in the intensities of two lights, or in the pitch of two tones, or in the heaviness of two weights. The ability to make fine sensory discriminations is an important factor in esthetic appreciation, in scientific investigation and in many daily activities.

The lower the threshold, the greater is the subject's sensory efficiency. But a low absolute threshold does not guarantee a low differential threshold. Discrimination of differences between two stimuli involves factors which are not present in the measurement of simple sensitivity. In discrimination the subject must make comparisons which involve judgment; and this is complicated by the fact that, in many tests, the stimuli are not present simultaneously when the judgment is made. When two sounds are given in succession, the comparison is made after the stimuli have been given, and both must be held in mind.

The tests described in this section are concerned with only two sense modalities, vision and hearing. There are two reasons for this. In the first place, these senses are of greatest importance in everyday

life. A man may suffer serious impairment of his sense of smell or taste without being very greatly handicapped. His sense of touch may also be diminished without much disturbance. But subnormal vision or hearing will make a great difference in a man's daily life, while loss of either of these senses is obviously a serious handicap. The second reason is that far more work has been done in vision and in hearing than in the other sense fields, and the tests are better standardized.

### 1. Visual Acuity

The determination of visual acuity is complicated by the susceptibility of the human eye to various defects. Visual acuity may be impaired by variation in the length and shape of the eyeball, of the lens, or of the cornea. These defects result in a failure of the eye to bring light rays sharply to a focus upon the retina.

There are many tests for acuity of vision, of which the most familiar is the Snellen Chart. Rows of letters, ranging in size from about ninety millimeters to five millimeters in height, appear on the chart and are to be read by the subject. The distance at which each row of letters should be read by the normal eye is printed on the chart. The measure of acuity of vision is expressed as a fraction, in which the numerator is the distance between the chart and the subject, and the denominator is the distance at which the row under examination can be read by the normal eye. If the subject is twenty feet from the chart, and can read the line of letters marked *twenty feet*, but none below that line, his vision is said to be 20/20, or normal. If he must approach to within twenty feet of the chart in order to read the thirty-foot line, his vision is 20/30, or about two-thirds of the normal. If he can read the fifteen-foot line when twenty feet away from the chart, his vision is 20/15, which is better than normal—at least in the sense that he can see objects clearly at a greater distance than can the average person.

The most common visual defects are hyperopia, myopia, and astigmatism. Hyperopia, or far-sightedness, is often accompanied by a feeling of eye-strain, by headaches, by inflammation of the eyes, and by other symptoms of discomfort. The far-sighted person has difficulty in reading at a normal distance, say fourteen or fifteen inches from the eye, but little difficulty in reading at a greater distance.

Myopia, or near-sightedness, is an exceedingly common visual defect. Like far-sightedness, it may be corrected by the use of properly fitted glasses. Collins (4) who examined 9,245 children from six to sixteen years of age in South Carolina, Maryland, Delaware and New York, found that 10 per cent. of these children had poor vision, 20/40 or less in one eye or in both, while 27 per cent. were moderately defective. Collins and Britten (5), working with 4,862 native white school boys from six to sixteen years of age, and with 6,479 male white industrial workers over eighteen years of age in various parts of the United States, found that the percentage of persons with normal vision increased with age up to eighteen years, after which it declined. The rate of decline was rapid after the age of forty-five. The percentage of persons with markedly defective vision (20/50 or less), increased steadily after six years of age. The rate of increase was more rapid during school ages than in the early years of industrial life. Levy (26) conducted an investigation among 4,021 school children, ages five to seventeen, in Syracuse, New York. Taking a record of 20/40 as the standard of poor vision, he found that 9.5 per cent. of the children fell into this class. This result agrees closely with that of Collins.

A third form of optical deficiency is astigmatism, which is caused by irregularities in the curvature of the cornea. This condition makes it impossible for the eye to focus the vertical and horizontal aspects of an object simultaneously. Kempf, Jarman, and Collins (22) examined 1,860 school children of both sexes, ages six to sixteen, in Washington, D. C. They found that 28 per cent. of these children were troubled with astigmatism. This defect may be corrected by suitable glasses.

While the Snellen Charts are useful as a quick method of detecting visual deficiencies, they possess certain limitations. In the first place, the test type is not sufficiently varied, so that it is possible for subjects to memorize some of the lines. In the second place, the arrangement of the letters has been criticized. Some of the lines are easier to read than others, while some are especially difficult by reason of the similarity of the letters. In the third place, the test does not reveal astigmatism accurately, a separate test being necessary for this purpose. Finally, it is questionable whether the reading of letters and nothing else is a sufficient test of visual acuity. All of these objections are met fairly well by another test, modeled after

# TESTS OF PHYSICAL AND SENSORY CAPACITY

27

the Snellen plan. This is a series of charts developed by Ewing (10), and approved by the Ophthalmic Section of the American Medical Association.

The Ewing Test consists of a series of charts on which are printed varieties of characters to be read by the subject. One group of charts contains characters intended for persons who are not familiar with letters. These characters are a ring, a square, a star, a flag, a fork and a chair. Another group of charts contains letters of graduated

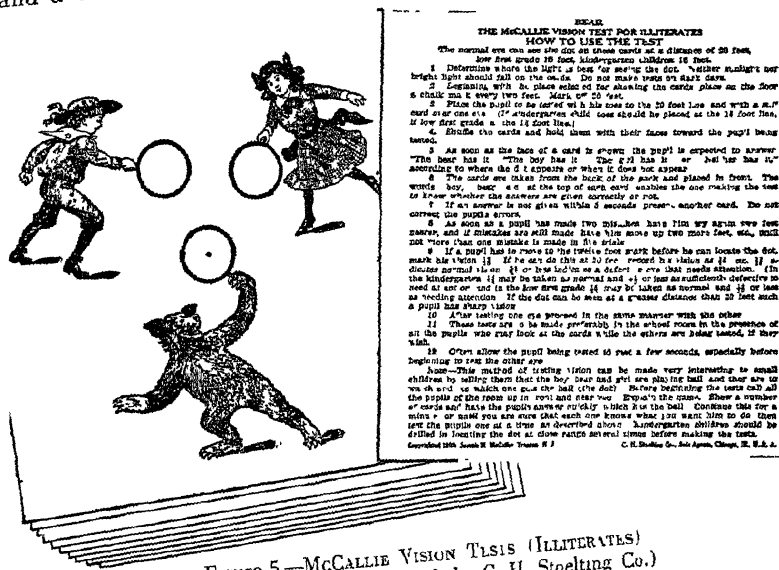


Figure 5.—McCallie Vision Tests (Illiterates)  
(Reproduced by courtesy of the C H Stoeckert Co.)

size, each size to be read at a given distance from the subject. A third group contains arrangements of black lines to form a cross, each arm of the cross consisting of three parallel lines, or a total of twelve lines. One of these twelve lines is broken; and the test consists in distinguishing the arm containing the broken line. The crosses are of various sizes, each marked with the distance at which the broken line should be detected by the normal eye. The crosses are also used as a test of astigmatism; since, to the astigmatic eye, the lines do not appear to be of equal depth. The test includes, in addition, two smaller cards to be used as tests of acuity in reading vision. Scoring of acuity of vision is identical with that used for the Snellen Test.

Another test which should be mentioned here is the McCallie Cards (36), especially the set used with illiterates. This test consists of a set of ten cards, each card showing a picture of a boy, a girl, and a bear. The boy and girl, as well as the bear, are holding hoops in the air; and the subject is required to tell which hoop contains a ball. This test is suitable for young children, since it can be treated as a game as well as a test.

### Test 11. Visual Acuity

*Apparatus:* The Snellen Chart, or the Ewing set of cards; preferably for children, the McCallie Cards. The Snellen Charts are sold by most optical supply houses. The Ewing set of cards is published by the C. V. Mosby Co., St. Louis, Missouri. The McCallie Cards are sold by the C. H. Stoelting Co., Chicago, Illinois.

*Method:* (a) Snellen Chart: Attach the chart to the wall at one end of the room, making sure that the light is satisfactory. The chart should be illuminated brightly and evenly, and the eye should not be stimulated directly by the source of illumination. Glare should be eliminated. Have S stand exactly twenty feet from the chart. Hold a sheet of cardboard in front of S's left eye, and ask him to read aloud the letters on the chart, beginning with the top line and working downward. Use the same procedure for the right eye. The score can be checked by placing S at the twenty-foot line, and then having him move first toward and then away from the chart, until the distance at which a given line can be read clearly is determined.

(b) The Ewing Cards: These cards are used in the same way as the Snellen Charts. The set includes letters, pictorial characters, and broken lines arranged in various forms. For careful testing, it is advisable to use at least one card of each type.

(c) McCallie Cards: Shuffle the cards and hold them before the subject, who is expected to tell whether the boy, the girl, or the bear has the ball. The child of kindergarten age should see the ball at a distance of sixteen feet; the first-grade child at eighteen feet; and older children at twenty feet.

*Record:* See instructions accompanying the test.

## 2. Color Blindness

Persons who are color blind will, in everyday life, frequently name colors correctly. They have learned that certain intense brightnesses—not colors—are called red, green, *etc.*, by persons about them. The differentiation is one of brightness, however, rather than of hue or quality. When the criterion of brightness fails, difficulty ensues and the colors are confused. Since this situation may arise at any time, it is important to discover cases of color blindness. Apart

from the inconvenience which may be caused by failure to differentiate colors, genuine danger may result.

Deficiency in ability to distinguish colors may take several forms, though the classification is not fully agreed upon (50). The most common form is red-green blindness, in which red and green are confused with each other and with other colors of the same brightness and saturation, especially the browns, the blues and the grays. Of relatively rare occurrence is blue-yellow blindness, which is usually pathological; and total color blindness, which is often accompanied by other forms of optical defect. There are many grades of incomplete color blindness, commonly called color weakness. In such cases the color qualities are the same as for the normal eye, but there is diminished sensitivity to color, and difficulty in comparing colors differing in brightness or in saturation.

Most investigators divide red-green color blindness into two major types, *protanopia* and *deutanopia*. The protanopes are the red-blind. The spectrum for them consists only of blues and yellows of various degrees of brightness and saturation. The red end of the spectrum is shortened, red appearing as black or dark gray. The deutanopes are the green-blind. For these individuals, the spectrum consists also only of blues and yellows, but it is not shortened at the red end, as it is for the protanopes. This classification is fairly distinct, but there are many cases that will not fall definitely into either of the two classes mentioned.

In testing for color blindness, colored woolen skeins have been popular for many years, the standard series being that of Holmgren. Methods of using these skeins have been varied, but all are concerned with the confusion of red with green. Other standard tests are the Nagel Card series, the Stilling Plates, and a later test modeled after the Stilling Series, i.e., the Ishihara Test. There are still other tests in use, but most of them require apparatus which is beyond the resources of most institutions. A comprehensive discussion of tests for color blindness will be found in Haupt (17). We have chosen three tests for description here. These are the Holmgren Woolens, the Nagel Cards, and the Ishihara Plates. These tests were selected because of their simplicity and ease of administration.

#### Test 12. Color Blindness

- (a) *The Holmgren Woolens*: Obtainable from any reputable optical supply house.

*Method:* Place the skeins, in irregular order, on a sheet of gray card-board. (1) Hand S the *large* green skein and ask him to select all of the skeins which resemble it in color. An exact match is not required. (2) If S shows hesitation, or if he selects gray, brown and red skeins as well as the green, give him the *large* rose skein, with directions as before. Color-blind subjects will often select blues and purples as matches and even greens and grays. (3) Finally, hand S the *large* red skein. Ordinarily there is little difficulty with this skein because of its saturation; but many color-blind subjects will select the greens and the dark blues.

(b) *The Nagel Cards:* A set of cards, in two sections, A and B. Obtainable from the C. H. Stoelting Co., Chicago, Illinois.

*Method:* Spread the sixteen cards of Section A on the table. S is instructed to make his responses by pointing with a pencil or with his finger. Ask the subject to point out the following cards, in the order given: (1) Cards with only red or reddish spots; (2) cards with red spots only; (3) cards with green spots only; (4) cards with gray spots only. Remove the cards and show S, one at a time, the four cards of Section B. Instruct him in each case to name the colors seen. Color-blind subjects, in answering questions 1 and 2, above, will select cards having brown and yellow-brown as well as red spots. In response to questions 3 and 4, the color-blind subject will usually confuse grays, greens and reds. Color-blind subjects will usually see only one color on each of the B cards: green on B1; red on B2 and B4; and gray or green on B3. If B2 and B4 are seen as single colors, or as red or green, the subject is definitely color blind.

(c) *The Ishihara Test:* A booklet containing sixteen color plates. Obtainable from the C. H. Stoelting Co., Chicago, Illinois, and from the Marietta Apparatus Co., Marietta, Ohio.

*Method:* Show S the pages of the booklet, one at a time, asking him to read aloud the numbers seen. Write down the numbers called. The colored plates are designed so that the normal eye can read the numbers clearly, while the color-blind eye will have difficulty, reading some of the numbers incorrectly and failing to read others. Compare the responses with the instructions given in the first two pages of the booklet. These instructions give the responses to be expected from normal persons, from the red-blind, the green-blind, and the totally color blind.

### 3. Auditory Acuity

All auditory tests should be given in a sound-proof room. If this is impossible, the examiner should at least arrange conditions so that extraneous sounds are reduced to a minimum. Probably the most important function of the human ear is the detection of spoken sounds. The best auditory test would, accordingly, be one in which

the subject is required to hear whispered and spoken speech. Such tests are rarely satisfactory, however, because it is difficult for the examiner to hold constant the intensity and intonation of his voice, even during the course of a single test. Nevertheless, at least one recent study (35), based upon sixty-one normal school students as subjects, finds a reasonably high reliability (average  $r$  for both ears = .78) for the whispered speech test, when trials by two different examiners are correlated. This is a higher reliability than is shown for the watch test in the same investigation, this result being .58 for the left ear and .64 for the right ear. An audiometer developed recently by the Western Electric Co., and used in an extensive survey among 4,419 grade-school children in Syracuse, New York (25), employs speech ranging from a loud conversational tone to a barely audible whisper. It is calibrated in terms of sensation units (11), and may be used for forty persons at a time by the employment of ear-phones. Fletcher (11) reports that it is possible with this apparatus to test 75 to 150 children per hour.

The watch has been widely used for testing auditory acuity, although it has certain obvious deficiencies. Unless the same watch or watches of similar make are used, it is impossible to compare results or to establish consistent norms. But when more precise methods are unavailable, the watch is of service, provided the conditions under which the tests are administered are constant from test to test.

The tuning fork is another well-known instrument for measuring auditory acuity. It is easier to standardize this instrument than to standardize watches, and standard forks are easily obtainable. One obvious objection to the tuning fork is that it does not present the variety of sounds which we are accustomed to hear in everyday life. It is probable, too, that musical tones carry farther than those sounds which we are accustomed to hear in everyday life. The same objection, of course, may be raised against the watch. Those familiar with the concert hall know that a delicate tone will often be heard in the most remote parts of the hall, while the speaking voice may be indistinct. Our inability to equate the sounds used in our tests with those which are met with in the school and in the street explains the continued use of the speech test, despite its manifest failings.

There are several instruments which measure auditory acuity fairly accurately. Perhaps the best of these is the audiometer devised by C. E. Seashore (41). This is a delicate instrument which, unfor-



tunately, is fairly expensive. The audiometer is a well-standardized instrument, but, like the watch and the tuning fork, is open to the objection that its stimuli are artificial. These stimuli consist of a series of clicks of graduated intensity, which are heard through telephone receivers held close to the ears.

### Test 13. Auditory Acuity

#### (a) *Whispered Speech Test:*

*Materials:* Rubber stoppers or cotton to be used as ear plugs.

*Method:* This test is best performed in a sound-proof room. If this is

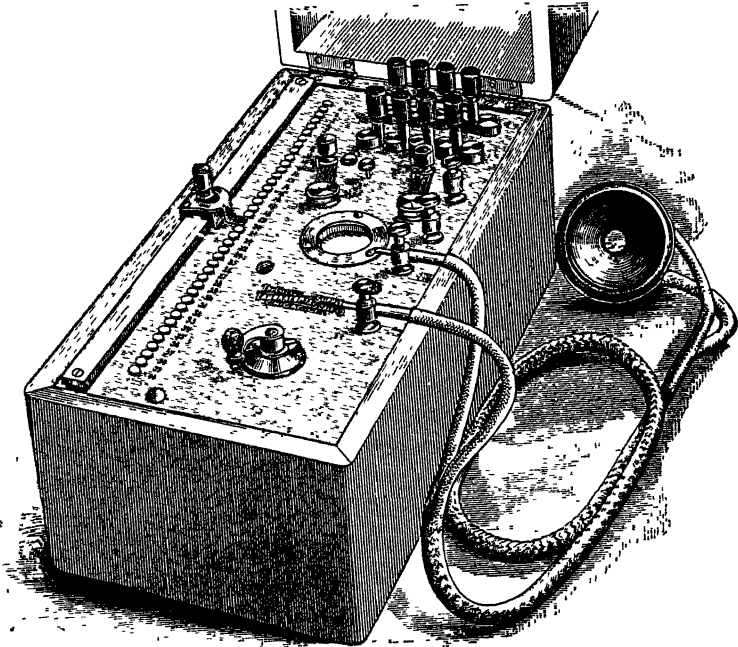


Figure 6.—SEASHORE AUDIOMETER  
(Reproduced by courtesy of the C. H. Stoelting Co.)

impossible, select a room as free as possible from sound. It is difficult to specify the exact range at which the subject should be required to hear the whispered sounds. Under reasonably good conditions, a suitable distance is probably forty-five feet. Have the subject sit with his right ear turned toward the experimenter. Plug the left ear, without causing discomfort. Use the list of numbers in Table VIII as stimuli.

Signal to S when a number is to be spoken. S should repeat the number, the responses being written down by E. Repeat the test for the left ear, using a different set of numbers. Most of the experiments cited in the preceding discussion have also tested both ears at once.

TABLE VIII  
TEST-NUMBERS FOR AUDITORY ACUITY  
(Whipple, after Andrews [53])

I	II	III	IV	V	VI	VII	VIII	IX	X
6	84	19	90	25	14	8	52	73	24
29	69	53	7	13	31	93	35	41	95
42	17	34	39	46	9	27	64	16	62
87	92	28	62	7	65	60	81	95	49
53	33	97	84	54	98	15	6	57	80
94	26	45	21	70	76	74	19	38	71
70	50	72	56	91	40	36	78	20	16
35	75	60	75	83	23	49	40	89	3
18	48	3	43	68	52	82	23	64	58
61	1	86	18	92	87	51	97	2	37

*Record:* The score for each of the three tests may be expressed as a percentage, *i.e.*, the number of items correctly heard divided by the total numbers of items in the list. If more than one series is given for each ear, take the percentage of the entire test, including all series. Since it is practically impossible to establish norms which will be generally applicable, each examiner should establish his own norms, based upon data derived from all of his subjects.

(b) *Watch Test:*

*Apparatus:* A watch which can be used for all tests, preferably a laboratory stop-watch; a yard stick; cotton or rubber ear plugs.

*Method:* Mark off on the floor a distance of thirty to forty yards, subdivided into feet or half-yards. S should not carry a watch during the test. Have S sit at one end of the distance marked off, his right ear turned toward the examiner, the other ear carefully plugged. Hold the watch in the palm of the hand, the face turned toward S. Begin with the watch close to S's ear, and move away quietly until he can no longer hear the ticking. Record this distance. When E judges that S is reaching his limit, the watch should be stopped; then started again several times until S answers correctly two out of three times when asked whether he hears the tick. Beginning at a distance too far for the subject to hear the tick, move the watch toward him until he can just hear the sound. Use the same checks as before, and record the maximum distance. Repeat until the two distances, forward and reverse, are consistent. Cover the watch with the hand from time to time, to make sure that S really hears the tick and is not imagining it. Make the test for the other ear and for both ears at once.

*Record:* The limen is obtained by averaging the distances recorded in both directions. Norms should be established from the records of all the subjects.

(c) *Tuning Fork Test:*

*Apparatus:* Blake's tuning fork, stop-watch. This tuning fork, which

has a vibration rate of 512 vs. per second, is obtainable from the C. H. Stoelting Co., Chicago, Illinois.

*Method:* E stands behind S, who is seated, with one ear plugged. Sound the tuning fork by squeezing the prongs together with the fingers and releasing them suddenly. Bring the fork close to the ear being tested. Instruct S, who is operating the stop-watch, to start the watch as soon as he begins to hear the sound of the fork, and to stop it as soon as the tone ceases to be heard. Test each ear five times, alternating from the right to the left ear.

*Record:* The record for each ear is the average time that the tone is heard in five trials. This average may be compared with the norm established empirically from the records of a number of subjects.

(d) *Seashore Audiometer Test:*

*Apparatus:* Seashore audiometer. This instrument is obtainable from the C. H. Stoelting Co., Chicago, Illinois.

*Method:* Detailed instructions accompany this instrument. These must be followed exactly if accurate results are to be obtained.

#### BIBLIOGRAPHY

1. ABERNETHY, E. M., "Correlations in Physical and Mental Growth," *Journal Educational Psychology*, 16:458-466, 1925.
2. BOAS, F., *The Mind of Primitive Man*, The Macmillan Company, New York, 1919.
3. BOVARD, J. F., AND COZENS, F. W., *Tests and Measurements in Physical Education*, W. B. Saunders Company, Philadelphia, 1930.
4. COLLINS, S. D., "The Eyesight of the School Child as Determined by the Snellen Test," *U. S. Public Health Reports*, U. S. Public Health Service, No. 48, 39:3013-3027, 1924.
5. COLLINS, S. D., AND BRITTEN, R. H., "Variation in Eyesight at Different Ages, as Determined by the Snellen Test," *U. S. Public Health Reports*, U. S. Public Health Service, No. 51, 39:3189-3194, 1924.
6. COLLINS, S. D., AND HOWE, E. C., "A Preliminary Selection of Tests of Fitness," *American Physical Education Review*, 29:563-571, 1924.
7. DEBUSK, B. W., "Height, Weight, Vital Capacity and Retardation," *Pedagogical Seminary*, 20:89-92, 1913.
8. DIXON, R. B., *The Racial History of Man*, Charles Scribner's Sons, New York, 1923.
9. DREYER, G., AND HANSON, G. F., *The Assessment of Physical Fitness*, Cassell & Company, London, 1920.
10. EWING, A. E., *Visual Test Types*, The C. V. Mosby Company, St. Louis, 1926.
11. FLETCHER, H., *New Methods and Apparatus for Testing the Acuity of Hearing*, Bell Telephone Laboratories, Reprint B-152-1:124, 1925.
12. FOWLER, E. P., AND FLETCHER, H., "Three Million Deafened School Children," *Journal American Medical Association*, No. 23, 87:1877-1882, 1926.

13. GARFIEL, E., "The Measurement of Motor Ability," *Archives Psychology*, No. 62, 9, 1923.
14. GATES, A. I., "The Nature and Educational Significance of Physical Status and of Mental, Physiological, Social and Emotional Maturity," *Journal Educational Psychology*, 15:329-358, 1924.
15. GITTINGS, I. E., "Correlation of Mental and Physical Traits in University of Arizona Freshmen Women," *American Physical Education Review*, 32:569-583, 1927.
16. GODDARD, H. H., "The Height and Weight of Feeble-minded Children in American Institutions," *Journal Nervous and Mental Diseases*, 39: 217-235, 1912.
17. HAUPT, I. A., "Tests for Color-Blindness: A Survey of the Literature, with Bibliography to 1928," *Journal General Psychology*, 3:222-267, 1930.
18. HULL, C. L., *Aptitude Testing*, World Book Company, Yonkers, New York, 1928.
19. HUNTINGTON, E., *The Character of Races*, Charles Scribner's Sons, New York, 1924.
20. JOHNSON, B. J., *Mental Growth of Children in Relation to Rate of Growth of Bodily Development*, E. P. Dutton & Company, New York, 1925.
21. KEMPF, G. A., AND COLLINS, S. D., "A Study of the Relation Between Mental and Physical Status of Children in Two Counties of Illinois," *U. S. Public Health Reports*, U. S. Public Health Service, No. 29, 44:1743-1784, 1929.
22. KEMPF, G. A., JARMAN, B. L., AND COLLINS, S. D., "A Special Study of the Vision of School Children," *U. S. Public Health Reports*, U. S. Public Health Service, No. 27, 43:1713-1739, 1928.
23. KRETSCHMER, E., *Physique and Character*, Harcourt, Brace & Company, New York, 1925.
24. KROEBER, A. L., *Anthropology*, Harcourt, Brace & Company, New York, 1923.
25. LAURER, F. A., "Hearing Survey Among a Group of Pupils of Syracuse Schools," *American Journal Public Health and the Nation's Health*, 18:1353-1360, 1928.
26. LEVY, H. H., "Vision Survey Among a Group of Pupils of Syracuse Schools," *American Journal Public Health and the Nation's Health*, 18:1273-1281, 1928.
27. Michigan State Council Meeting of Physical Education, "Physical Education in the State of Michigan: Pulse-rate for Physical Fitness," *American Physical Education Review*, 25:138-139, 1920.
28. MONAHAN, J. E., AND HOLLINGWORTH, L. S., "Neuro-muscular Capacity of Children Who Test Above 135 I.Q. (Stanford-Binet)," *Journal Educational Psychology*, 18:88-96, 1927.
29. MURDOCH, K., AND SULLIVAN, L. R., "A Contribution to the Study of

- Mental and Physical Measurements in Normal Children," *American Physical Education Review*, 28:209-215, 270-280, 328-330, 1923.
30. NACCARATI, S., "Morphologic Aspect of Intelligence," *Archives Psychology*, No. 45, 1921.
  31. NACCARATI, S., AND GUINZBERG, R. L., "Hormones and Intelligence," *Journal Applied Psychology*, 6:221-234, 1922.
  32. PATERSON, D. G., *Physique and Intellect*, The Century Company, New York, 1930.
  33. PEARSON, K., "On the Relation of Intelligence to Size and Shape of Head and to Other Physical and Mental Characters," *Biometrika*, 5: 105-146, 1906-1907.
  34. PEATMAN, J. G., "A Study of Factors Measured by the Thorndike Intelligence Examination for High School Graduates," *Archives Psychology*, No. 128, 1931.
  35. PETERSON, H. A., AND KUDERNA, J. G., "Reliability of School Tests of Auditory Acuity," *Journal Educational Psychology*, 15:145-156, 1924.
  36. PYLE, W. H., *The Examination of School Children*, The Macmillan Company, New York, 1913.
  37. RADOSAVLJEVICH, P. R., "Prof. Boas' New Theory of the Form of the Head," *American Anthropologist*, N.S. 13:394-436, 1911.
  38. REID, R. W., AND MULLIGAN, J. H., "Relation of Cranial Capacity to Intelligence," *Journal Royal Anthropological Institute*, 53:322-331, 1923.
  39. ROGERS, F. R., *Physical Capacity Tests in the Administration of Physical Education*, Teachers College, Columbia University, Contributions to Education, 173, 1925.
  40. RUDISILL, E. S., "Correlations Between Physical and Motor Capacity and Intelligence," *School and Society*, 18:178-179, 1923.
  41. SEASHORE, C. E., "An Audiometer," *University of Iowa Studies in Psychology*, 2:148-163, 1899.
  42. SHELDON, W. H., "Morphologic Types and Mental Ability," *Journal Personnel Research*, 5:447-451, 1926-1927.
  43. SHELDON, W. H., "Social Traits and Morphologic Types," *Personnel Journal*, 6:47-55; 1927-1928.
  44. SHERMAN, E. B., *An Experimental Investigation Concerning Possible Correlation Between Certain Head Measurements and University Grades*, Thesis, University of Wisconsin Library, 1923 (see Hull [18]).
  45. SMITH, H. L., AND WRIGHT, W. W., *Tests and Measurements*, Silver, Burdett & Company, New York, 1928.
  46. SOMMERVILLE, R. C., "Physical, Motor and Sensory Traits," *Archives Psychology*, No. 75, 12, 1924.
  47. STALNAKER, E. M., "A Comparison of Certain Mental and Physical Measurements of School Children and College Students," *Journal Comparative Psychology*, 3:181-239, 1923.

48. STERLING, E. B., AND BELL, E., "Hearing of School Children as Measured by the Audiometer and as Related to School Work," *U. S. Public Health Reports*, U. S. Public Health Service, No. 20, 45:1117-1130, 1930.
49. STOCKARD, C. R., *Physical Basis of Personality*, W. W. Norton and Company, New York, 1931.
50. TERMAN, S. W., "A New Classification of the Red-Green Color-Blind," *American Journal Psychology*, 41:237-251, 1929.
51. TREDGOLD, A. F., *Mental Deficiency*, William Wood & Company, New York, 1929.
52. TURNER, A. H., "The Vital Capacity of College Women," *American Physical Education Review*, 32:593-606, 1927.
53. WHIPPLE, G. M., *Manual of Mental and Physical Tests, Simpler Processes*, Warwick and York, Baltimore, 1914.
54. WILLIAMS, J. F., *Principles of Physical Education*, W. B. Saunders Company, Philadelphia, 1927.
55. WILSON, M. G., AND EDWARDS, D. J., "Diagnostic Value of Determining Vital Capacity of Lungs of Children," *Journal American Medical Association*, 78:1107-1110, 1922.
56. WISSLER, C., "Sex Differences in Growth of the Head," *School and Society*, 25:143-146, 1927.
57. WOOD, T. D., *Personal Health Standard and Scale*, Bureau of Publications, Teachers College, Columbia University, New York, 1925.
58. WOOLLEY, H. T., AND FISHER, C. R., "Mental and Physical Measurements of Working Children," *Psychological Monographs*, No. 1, 18, 1914-1915.

## CHAPTER II

### TESTS OF MOTOR ABILITY AND MECHANICAL APTITUDE

TESTS of motor ability and of mechanical skill have long been of interest to psychologists and are becoming increasingly valuable to educators and teachers. Motor tests were first studied extensively by psychologists interested in the problem of individual differences, and in the relation of physical and motor traits to estimates of mental ability. Such tests have been employed extensively in psychological research, especially in clinical studies of atypical children, in industrial selection, and in investigations of sex and race differences. An increasingly important rôle, also, is being assigned by modern educators to training in physical coördination and in the development of manual skill. Hence, in many schools, extensive use is now made of motor and mechanical tests.

The present chapter deals with some representative tests of the speed of movement, muscular coördination, precision and agility. Tests of mechanical and constructive aptitudes are also discussed. While these latter tests are not essentially motor (see p. 55) they usually involve manual skill and hand-eye coördination, and hence seem to group naturally with motor tests.

#### TESTS OF SPEED OF MOVEMENT AND MUSCULAR CONTROL

##### 1. Tapping

The tapping test was designed to measure speed of voluntary movement, without ordinarily demanding much precision or coördination. In one form or another tapping has been used for many purposes; and methods and techniques have varied almost as widely as the uses to which the test has been put. Several investigators (37) have reported sex differences in the speed of tapping, the general conclusion being that men are faster than women and boys faster than girls. As a measure of motor speed the principal value of the tapping test would seem to lie in its relationship to other functions

and in its utility in indicating changes in the condition of the subject. The tapping test, for example, has been used as a measure of the effects of caffeine, tobacco, alcohol, change of diet and loss of sleep. Typical of such studies is Hull's investigation of the effects of tobacco, and Hollingworth's study of the effects of caffeine and alcohol. Hull (23), using nineteen male adults as subjects, found a slight loss in the speed of tapping in habitual smokers after smoking, and a slight increase in tapping speed in non-smokers after smoking. These effects disappeared at the end of about an hour. Hollingworth (20) studied the effects of caffeine upon tests requiring speed of movement, coördination and steadiness. A small but definite increase in speed of movement (tapping) of about 4 per cent. appeared as a result of a dose of caffeine equivalent to one to three cups of coffee. This increase was not followed by any subsequent loss in speed. Hollingworth's (22) investigation of the effects

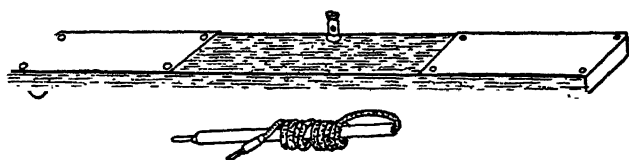


Figure 7.—TAPPING TEST

(Reproduced by courtesy of the Marietta Apparatus Co.)

of alcohol also included the tapping test. His records for six subjects indicated a loss in speed after alcohol consumption ranging from 7 per cent. to 14 per cent., the degree of loss depending directly upon the quantity of alcohol drunk.

A few experiments have investigated the possible uses of the tapping test as an index of vocational aptitude. A good example is Link's study (30) of the relationship between tapping and efficiency in factory work as measured by actual production. Link's subjects were seventy-three girls in a firearms factory. In a group of fifty-two girls employed to inspect shells, the correlation between production and tapping was only .14; but in a group of twenty-one girls engaged in fitting shells into a gauge, the correlation was .52. The job of gauging shells is largely motor, while inspecting is mainly visual, and this difference probably accounts for the rise in correlation. When we remember that correlations between mental tests and school



success are seldom higher than .50, Link's result for gaugers is rather striking. Another interesting study in this field is that of Kitson (27), in which the occupations studied, *viz.*, typing and piano playing, have an apparently direct relationship to the tapping activity. Kitson's subjects were twenty-five women pianists who had studied from four to eleven years, and twenty-five women typists. These two groups were matched against control groups of the same sex and ages but untrained in the given vocations. The differences in average score between the pianists or the typists and the other groups were small and unreliable. There was evidence, however, that those of the highest ability in typing and in piano playing, and perhaps the failures also, could be "spotted" by a tapping test. For the intermediate groups the results were inconclusive.

The degree of relationship between tapping and mental ability apparently depends largely upon the ages of the subjects tested. Burt (8) and Abelson (1), for instance, report correlations between tapping and various mental tests which range from .28 to .65. Burt's group consisted of forty-three boys 12½ to 13½ years old; Abelson's group of eighty-eight girls and forty-three boys, all backward mentally, and ten, eleven and twelve years old. Sommerville (43), on the other hand, found a correlation of —.09 between the Thorndike Intelligence Test and speed of tapping in a group of 105 college students. If age variability is rendered constant, this correlation becomes zero. Garfiel (15), in a group of thirty-two freshmen women, reports a correlation between tapping and Army Alpha of —.12. It seems unreasonable to expect that a simple function like tapping would be intimately related to measures of intellectual activity in adults of college level. In children, however, the situation is quite different, for tapping, as well as other tests which are simple for the adult, constitute real tasks for them.

The tapping test has been utilized in the development of many teams or batteries designed to measure motor ability, motor skills and mechanical aptitude. Batteries of this kind have been developed by Garfiel (15), Seashore (41), Cowdery (11), and Paterson, Elliott, *et al.* (36). Superficially, the tapping test taken by itself does not appear to have given results commensurate with the labor expended upon it. But two facts should be considered in evaluating the test. First, the test as ordinarily used is very short. While it is

desirable to use short tests to predict aptitudes, it is improbable that a short test of a very simple function will forecast efficiency in a complex activity. Second, tapping is primarily a measure of the speed of finger and arm movements and hence can hardly be expected to give high correlations with activities which demand much more than speed, and involve larger muscle groups. Nevertheless, while tapping usually shows low correlations with criteria of motor ability, in combination with other motor tests it is frequently useful.

It is difficult to find conclusive data on the reliability of the tapping test, because of the diversity of methods employed. Gates (16), correlating various practice periods of the tapping test, has reported reliabilities ranging from .62 to .91. Lanier (29), working with a form of tapping which required his group of 104 subjects to place a dot in each of a number of small squares, found the reliability of the test with five-millimeter squares to be .63. Muscio (34), in a group of eighty-eight subjects consisting of both sexes and varying widely in age, correlated the combined scores of the first and fourth trials in a tapping test with the second and third trials. This gave a reliability of .86; correlation of the twenty-first and the twenty-fourth with the twenty-second and twenty-third trials gave a reliability of .92. These results indicate that the reliability of the tapping test increases with practice, and is greatest after the limits of improvement have been reached.

Of the many methods used to investigate speed of tapping, two will be described here. These are the tapping plate and the telegraph key methods.

#### Test 1. Tapping

(a) *Apparatus:* Dunlap alternate tapping plate (13); Veeder counter (Ream); tapping stylus; dry-cell batteries; stop-watch; knife switch. (See Figure 7.)

*Obtainable from:* The C. H. Stoelting Co., Chicago, Illinois, or from the Marietta Apparatus Co., Marietta, Ohio.

*Method:* The tapping plate, the counter, the stylus, and switch are connected in series. Instruct S to hold the stylus in his right hand and at the signal *go*, to tap on one of the plates as rapidly as possible. A fairly standard movement is secured by having S rest his elbow on the table, and use both wrist and elbow joints in tapping. At the *go* signal, E throws in the switch, at the same time starting the watch. At the end of sixty seconds E throws out the switch, stops S and reads the counter. Repeat the test for the left hand. A variation of this method is to have

S tap on both plates alternately, in which case two counters are needed.  
*Record:* The score is the total number of taps recorded in a trial of sixty seconds.

*Alternative Method:* Instead of the counters, a kymograph drum, which gives a graphic record, may be used (49). When a Jacquet chronograph is used the record may be taken for any part of the sixty-second period.

(b) *Apparatus:* The same as in (a) except that a telegraph key is used instead of the tapping plate and stylus.

*Method:* The same as in (a).

(c) *Note:* A good typewriter may be used, a single key being tapped as rapidly as possible. The speed, however, is limited by the flexibility of the machine. Another method is to provide S with four sheets of paper, spread out. At the signal *go*, S taps on the first sheet with a pencil; after fifteen seconds the signal is given and S shifts to the second sheet, and so on through the third and fourth sheets (14). A variation of this method consists in having S tap with a pencil in each of a succession of printed squares (29). This task, however, requires precision as well as speed.

*Norms:* Tapping norms for children are given in Tables IX and X. Table IX is taken from Smedley (49), who used a tapping board. Table X was compiled by Bronner, Healy, *et al.* (7), from results secured with a dot and square test.<sup>1</sup> In the Columbia laboratory the average number of taps (using tapping board) was 196.91 (S.D. = 26.83) in five trials of thirty seconds each. These results were obtained with a group of sixty-eight men and women, graduate students in psychology.

TABLE IX  
 TAPPING NORMS BY SEX AND AGE (TAPPING BOARD)  
 (From Whipple, after Smedley [49])

Age	Number Tested	Boys		Number Tested	Girls	
		Taps in 30 Seconds Right Hand	Taps in 30 Seconds Left Hand		Taps in 30 Seconds Right Hand	Taps in 30 Seconds Left Hand
8 . . . .	31	117	117	31	146	117
9 . . . .	60	151	127	44	149	118
10 . . . . .	47	161	132	48	157	129
11 . . . .	49	162	111	48	169	139
12 . . . .	41	170	115	50	169	140
13 . . . .	50	184	156	45	178	153
14 . . . . .	40	184	155	67	181	157
15 . . . . .	37	191	169	48	181	159
16 . . . .	21	196	170	50	188	167
17 . . . .	13	196	174	40	184	162
18 . . . .	3	197	183	24	193	169

<sup>1</sup> In this test, S is given a blank ruled in squares. The task is to make a tap in each square without touching the lines.

TABLE X  
TAPPING NORMS BY SEX AND AGE (DOT AND SQUARE TEST)  
1,205 Males                      812 Females  
(Score is number of dots made in 30 seconds)  
(Bronner, Healy, *et al* [7])

Sex	Age	8	9	10	11	12	13	14	15	16	16+
M.	75 Percentile	54	60	65	73	76	81	87	90	98	99
	50 Percentile	51	53	58	63	67	74	80	81	87	91
	25 Percentile	45	47	51	55	60	66	71	73	78	79
F.	75 Percentile			71	78	82	92	94	97	100	103
	50 Percentile			63	69	73	81	85	87	91	95
	25 Percentile			54	63	66	72	77	80	82	87

## 2. Steadiness

This is a standard test, designed to measure control of muscular tremor. It has been used for much the same purpose as the tapping test, although the two tests as shown by their low intercorrelation measure different performances. Griffitts (18), who correlated the steadiness and the tapping tests given to a group of sixty male college students, obtained a correlation coefficient of .04 for the right hand and of .20 for the left hand. Garfel (15) found the very low correlation of .08 between steadiness and tapping in a group of fifty girls. These low relationships indicate the possibility of combining the steadiness test with other motor speed tests, when the purpose is to form a battery which will give a high correlation with some activity involving both speed and steadiness.

Both Hull and Hollingworth, in the experiments referred to on p. 39, employed the steadiness test in studying the influence of drugs upon efficiency. According to Hull, tobacco smoking produces a marked tremor of the hand, which is about as pronounced for habitual smokers as for non-smokers. Recovery is practically complete in an hour and a half after smoking. Hollingworth found that consumption of alcohol produced a marked decrease in steadiness, ranging from a 68 per cent. loss after three to four bottles of beer, to a 370 per cent. loss after six to nine bottles. Recovery was relatively slow. Hollingworth's study of caffeine indicated that one to four grains of caffeine (the average cup of coffee with cream contains 2.5 grains of caffeine) produced a slight decrease in steadiness which persisted for several hours. The loss of steadiness was much more noticeable when the dose of caffeine was six grains or more.

The correlation between steadiness and intelligence test scores is quite low. Garfiel (15) found a correlation of .27 between steadiness and Army Alpha, and Sommerville (43) a correlation of  $-.14$  between steadiness and the Thorndike Intelligence Examination.

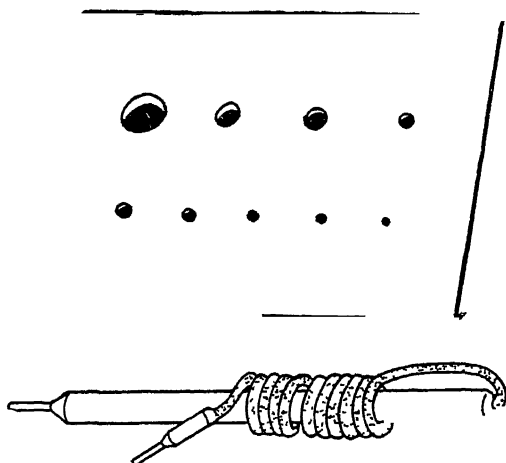


Figure 8.—STEADINESS TEST

(Reproduced by courtesy of the Marietta Apparatus Co.)

According to Griffitts (18), the reliability of the steadiness test was .71 for a group of sixty college students.

## Test 2. Steadiness

*Apparatus:* Steadiness tester, with nine holes; metallic needle; electric counter; dry-cell batteries; knife switch; stop-watch. (See Figure 8.)

*Obtainable from:* The C. H. Stoelting Co., Chicago, Illinois, or from the Marietta Apparatus Co., Marietta, Ohio.

*Method:* Connect the steadiness tester, the needle, the counter, the switch, and the batteries, in series. Have S sit in front of the apparatus and hold the needle in his right hand. Instruct him to insert the needle into the largest hole and hold it there trying not to touch the sides. Throw in the switch when the go signal is given, and start the stop-watch. At the end of fifteen seconds throw out the switch and note on the counter the number of contacts made by S. Test now for the other holes in order, from largest to smallest. Be careful that S does not hold the needle against the plate continuously. In taking this test the arm and hand should not be supported, and the forearm and upper arm should form an angle of about  $100^\circ$ . Allow thirty seconds' rest between trials.

*Alternative Method:* A kymograph drum may be used to record the contacts instead of a counter. This gives a permanent graphic record.

*Record:* The contacts are recorded by the counter, or on the kymograph drum. Tabulate the number of contacts made in each of the holes. Johnson (25) used as the score the smallest hole in which not more than twelve contacts were made. Dewey, Child and Ruml (12) assign numerical values to each of the nine holes. These values range from 1 to 105 by steps of 13, hole 1 having a value of 105, hole 2 a value of 92, hole 3 a value of 79, *etc.* An individual's score is the numerical value of the hole in which he had made *less* than twelve contacts in fifteen seconds, plus the actual number of contacts made in that hole. Thus, five contacts in hole 3 is a score of 84.

*Norms:* Table XI gives norms taken from Dewey, Child and Ruml in terms of the method of scoring described above.

TABLE XI  
STEADINESS NORMS  
Right-hand Score  
(Dewey, Child and Ruml [12])

Boys					
Age	9 0-9 9	10 0-10 9	11 0-11 9	12 0-12 9	13 0-13 9
Mean	82 8	72 8	72 2	59 9	62 3
S.D.	12 1	17 2	15 5	17 0	16 8
Girls					
Age	9 0-9 9	10 0-10 9	11 0-11 9	12 0-12 9	13 0-13 9
Mean	67 7	69 3	66 2	57 7	47 4
S.D.	18 3	16 3	17 4	17 4	14 5

Two other tests of steadiness may be mentioned briefly. These are the *taxiameter*, designed by Miles (33), and used to measure large-muscle control and body-sway; and the *automatograph* often used in the psychological laboratory to measure involuntary tremor of the arm (45).

### 3. A Scale of Motor Ability Tests

This scale was devised by Brace (6) and standardized upon 775 boys and girls, ages eight to nineteen, and 523 college women. It consists of twenty tests of body balance, descriptions and photographs of which are given in Brace's monograph. In a group of forty-four women, a retest after a year correlated .66 with the first test; in a group of 106 children, aged thirteen, fourteen and fifteen, the retest correlation was .82. Two criteria of validity were used by Brace: ratings for motor ability and scores made in a series of athletic

events. In a group of boys and girls the correlation of the scale with the ratings was .58; for boys alone, the correlation of the scale with scores in athletic events was .80. These results show the scale to be fairly satisfactory as an indicator of general motor control.

#### 4. Body-balancing

This test was used by Paterson, Elliott, *et al.* (36), in the development of the Minnesota Mechanical Ability Tests. A three-inch cube of wood is attached to the middle of a board one inch thick, two feet long, and five inches wide. Placing the ball of one foot upon the center of the block, S is required to balance himself on one leg. Two trials are taken fifteen minutes apart. The reliability of this test for 217 seventh- and eighth-grade boys was .56.

Another form of balancing test has been reported by Hertzberg (19). A "two by four" board, ten feet long, is set on its narrow edge and supported by braces. The child is required to walk along this edge and the distance at which he is forced to step off is recorded. The correlation between this test and Stanford-Binet mental age was .15 in a group of forty-six children of both sexes, when variability in age was held constant. If variability in M.A. is partialled out, the correlation of the test with chronological age is .41. This indicates that control of equilibrium is more closely related to physical maturity than to a measure of general intelligence.

#### 5. Scales Tests

This test is described by Griffiths (18). Two household scales are placed before S. The dial of the scale at the left faces S, and the dial of the scale at the right faces the experimenter and is concealed from S. S is required to push down the scale on his left until the dial registers one-half pound. Simultaneously, he pushes down the scale on his right until, in his opinion, he is exerting the same pressure with the right hand as with the left. This procedure is repeated with 1, 2, 3, 4, 5, 6, 8 and 10 pounds as "standard" pressures. Four trials are given on each standard and the discrepancy between right- and left-hand pressures recorded in pounds. In a group of sixty college men the reliability of this test was .72.

### TESTS OF MOTOR COÖRDINATION AND PRECISION

The tests described in this section require a somewhat greater degree of coördination and precision of movement than do tests of speed and muscular control.

### 1. Coördination (three-hole test)

This is a well-known test which has been often investigated along with steadiness, tapping and other standard motor tests. It usually consists of a triangular board tilted back at an angle of  $45^\circ$  from horizontal. (See Figure 9.) Three holes lined with brass are set in the three angles of the triangle. The subject is required to insert a stylus into each of these holes in succession, working as rapidly as possible. Each time the stylus is inserted into a hole a contact is registered on an electric or automatic counter. Carothers (9) has reported the reliability of this test to be .66 in a group of forty-five Barnard freshmen; and Garfiel (15) obtained a reliability coefficient of .50 in a group of fifty Barnard sophomores. Carothers reported low intercorrelations between the three-hole test and nineteen rela-

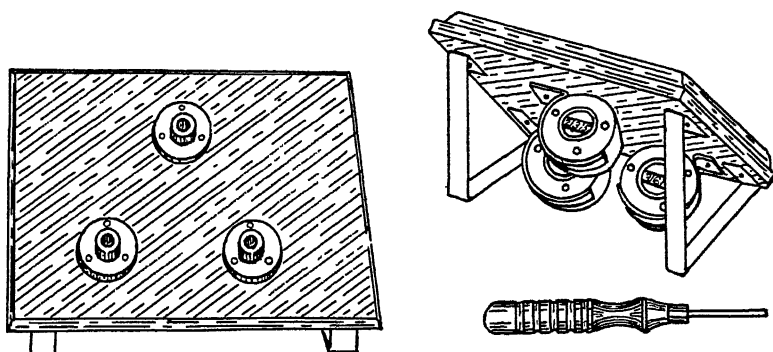


Figure 9.—COORDINATION TEST (THREE HOLE)  
(Reproduced by courtesy of the Marietta Apparatus Co.)

tively simple mental and physical tests given to a group of 100 women. The correlation of three-hole coördination and tapping, however, was .48, which indicates a fairly good relationship between these two activities. Hollingworth (21) found, in a group of thirteen subjects, that the correlation of the three-hole test with other tests varied with the stage of practice reached. The correlation with tapping, for example, rose from —.25 for the first trial to .39 for the two hundred fifth trial.

Hollingworth, in his studies of the effects of drugs upon performance (20, 22), reported a loss in coördination, as measured by the three-hole test, which varied directly with the quantity of alcohol consumed. Small doses of caffeine, however, produced a slight improvement in coördination; while larger doses, equivalent to two or



three cups of coffee (five to six grains), resulted in short initial improvement, followed by an average loss, after several hours, of 2 to 3 per cent. The loss in efficiency varied inversely with the body weight of the subject, the heaviest subjects being the least affected. The effect of caffeine was greatest when taken on an empty stomach.

Griffitts (18) has tested coördination by means of the steadiness test described on p. 44. A metronome was used to control the speed of movement. At the first stroke, S inserted the stylus into the first hole; at the second stroke he withdrew the stylus and let his hand drop to the table. At the third stroke he inserted the stylus into the second hole, and so on. The score was the number of holes into which S inserted the stylus without touching the sides. In another variation of the three-hole test, Bickersteth (4) used a brass plate, twenty-three centimeters square, containing twenty-four holes arranged in a circle. Procedure was similar to the three-hole coördination test, a stylus being inserted into each of the twenty-four holes in order.

### Test 3. Three-hole Coördination

*Apparatus:* Three-hole coördination board; metal stylus; dry-cell batteries; electric or automatic counters; stop-watch. (See Figure 9.)

*Obtainable from:* The C. H. Stoelting Company, Chicago, Illinois.

*Method:* S should be seated directly facing the board, with the stylus in his right hand and his elbow resting on the table. Instruct S to insert the stylus into each hole as rapidly as possible, taking them in order. Give five trials of sixty seconds each and average the records for the final score.

A variation in scoring is to take the time for 100 insertions of the stylus. By using three counters, one for each hole, it is possible to record the insertions for each hole separately. This is sometimes desirable, for some subjects, in their eagerness to achieve a high score, will often miss a hole and go on to the next.

*Norms:* Baldwin and Stecher (3) report the following results for 105 boys and girls:

Age	Insertions (per minute)
2 . . . . .	15 6
3 . . . . .	20 1
4 . . . . .	27 6
5 . . . . .	35 0
6 . . . . .	38 4

According to Carothers (9), the average number of insertions per minute for 200 Barnard freshmen was 83.4.

## 2. Aiming

Although there is little uniformity in the method of conducting

tests of aiming, all such tests are alike in requiring a rather high degree of precision of movement. Several aiming tests will be described in this section.

(a) A test blank containing ten crosses irregularly arranged has been devised by Whipple (49). This blank, shown in Figure 10, is attached to a wall and the subject is required to stand before the blank and to strike with a pencil point at the intersections of the crosses. The speed of thrusting is usually regulated by a metronome. The test is scored in terms of the amount by which S misses the intersections. With slight modifications, this test was used by Griffiths (18) in a group of sixty college students. The reliability of the test was .83 for the right hand and .74 for the left hand. This aiming test bears some relation to Griffiths' adaptation of the three-hole co-ordination test, the correlation of the two tests being .44. The correlation of aiming with steadiness was .39; with tapping, .00; and with MacQuarrie's dotting test, .11.

(b) In studying hand-eye coördination, Paterson, Elliott, *et al.* (36), arranged a target consisting of twenty-five concentric circles. The smallest circle had a radius of two millimeters, and each successive circle increased in radius by two millimeters. S was required to hold

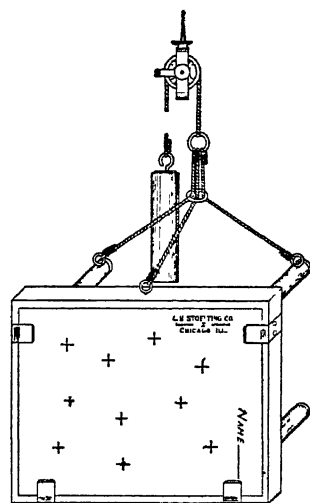


Figure 10.—AIMING OR  
TARGET TEST

(Reproduced by courtesy of the  
C. H. Stoelting Co.)

a pencil level with his eyes and to try to hit the center of the target ten times. This task was repeated three times, a fresh sheet being used for each trial. The score for each trial is the average distance of the ten dots from the center. When given to 217 boys in the seventh and eighth grades, the reliability of this test was only .24, which indicates a large chance element in the scores made.

(c) An interesting study of individual differences in the aiming test is reported by Blackburn (5), who modified Whipple's test by using circles instead of crosses. The subjects, 128 army recruits, were instructed to spear the centers of the circles with a metal pointer. The circles were connected by lines to indicate the order in

which they were to be struck. Blackburn contends that even in such simple tests as this, one should not rely entirely upon objective scoring methods. He believes that due consideration must be given to the way in which S performs the task, and to such factors as natural speed, sense of rhythm, *etc.*

(d) Another form of aiming test consists in throwing objects at targets. Such a test was used by Garfiel (15), who set up a target consisting of a bull's eye surrounded by three concentric circles. The subject was required to throw a rubber ball at the bull's eye from a distance of twelve feet. Carver (10) had eight subjects throw darts at a target, before and after smoking, and found that tobacco caused a slight loss in accuracy. Johnson (24) has used the dart-throwing test in a study of fifteen inmates in the Bedford Hills reformatory for women. This group was divided into three sub-groups of five each, the first consisting of the more intelligent inmates, the second of the somewhat less intelligent individuals, and the third of the feeble-minded. The most intelligent group had the highest final average score in dart-throwing and the greatest initial ability, the middle group ranked second, and the feeble-minded group lowest. The differences, however, were not significant. The learning curves for the two extreme groups were more irregular than that of the middle group.

(e) A test which consisted of throwing rings over a post has been used by Goodenough and Brian (17). Twenty four-year-old children were the subjects. Rope rings were thrown at a post from a distance of ten feet, twenty throws per day for fifty days constituting the experiment. Reliability coefficients were .59 for one day's performance, and .93 for the total performance of fifty days.

### 3. Tracing

An extensive experiment in tracing has been conducted by Wellman (47), who used the tracing board described by Whipple (49); and also a paper and pencil tracing path, in which the path was a duplicate of the Whipple tracing board. The paper test was employed instead of the board because it furnished a permanent record, and gave the experimenter ample opportunity to observe the child's mode of attack. The reliability of the tracing board was .82 in a group of fifty-four children from three to six years of age, while the reliability of the tracing path was .97 for ninety-four children from three to six years of age. The correlation between tracing path scores and

chronological age was .81 for boys and .82 for girls, indicating a decided relationship between age and this type of motor coördination. With variability in age held constant, the correlation between the tracing path and Stanford-Binet I.Q. was .09 for boys and .22 for girls.

#### Test 4. Tracing

*Apparatus:* Tracing board, shown in Figure 11; metal stylus; dry-cell batteries, telegraph sounder.

*Obtainable from:* The C. H. Stoelting Co., Chicago, Illinois, or from the Marietta Apparatus Co., Marietta, Ohio.

*Method:* Have S sit with the tracing board in front of him so that in tracing his hand will move toward his body in a median plane. Instruct S to place the tip of the stylus at the farther end of the groove, and

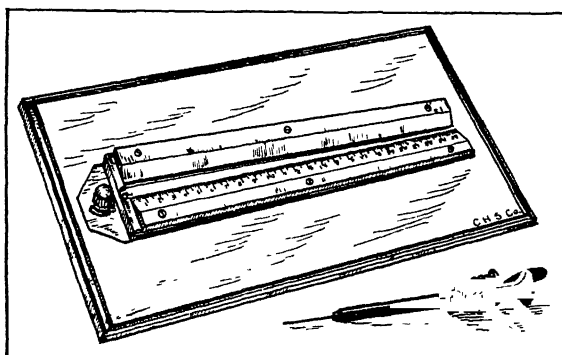


Figure 11.—TRACING BOARD

(Reproduced by courtesy of the C. H. Stoelting Co.)

to draw the stylus toward himself without touching either of the metal side strips. A full arm movement must be used. Give a few preliminary trials to adjust the speed, so that the board will be traversed in about ten seconds. Then have S take five trials with each hand, alternating from right hand to left. Stop S when the telegraph sounder indicates a contact. Note the point on the scale at which the contact was made.

*Record:* The score is the point on the scale at which contact was made.

*Alternate Method:* Another method is to require S to traverse the whole board and to note the number of contacts made. The points at which contacts occur may also be recorded. This last is easier if Wellman's tracing path is used, because a permanent tracing is secured, which may be scored at leisure.

#### 4. Miscellaneous Motor Tests<sup>1</sup>

(a) *Koerth Pursuit Apparatus:* This test was designed to measure

<sup>1</sup> Sold by the C. H. Stoelting Co., Chicago, Illinois.

hand-eye coördination (28). It consists essentially of a wooden disk upon which is a brass target, 1.9 centimeters in diameter, sunk flush with the surface of the disk. The disk is mounted on a phonograph, the speed of which can be varied and regulated. S, holding a pointer in his hand, tries to follow the revolutions of the disk while keeping the pointer on the target. The time during which S can keep contact with the target is recorded in tenths of a second. Percentile norms for 126 subjects are given by Koerth in graphic form. This test is part of the Stanford Motor Skills Unit (41), and its reliability for fifty university men is reported by Seashore (42) to be .93.

(b) *Miles Pursuitmeter* (32): The pursuitmeter requires the subject to follow the oscillations of a needle with a pointer. The needle oscillates at a different rate every five seconds, and the experiment lasts five minutes. Errors are recorded upon a revolving drum. A different form of pursuitmeter is described by Renshaw and Weiss (39). This test consists of a moving electrode with which S must keep contact, using a special hammer. The movements of the electrode change periodically. An analysis of the nature of the pursuit task is reported by Renshaw and Postle (40) and by Renshaw (38).

(c) *O'Connor Finger Dexterity Test*: This is one of a series of tests devised by O'Connor (35) and used by him in vocational selection. The subject is seated facing a board in which 100 holes have been drilled. A tray containing 310 pins is placed on S's right or left, depending upon whether he is right- or left-handed. The task consists in placing three pins in each of the 100 holes as rapidly as possible. The score is usually in terms of time. While this test is intended primarily for women, O'Connor has published norms for both men and women (Appendix B in O'Connor [35]). Norms for boys and girls, ages seven to fifteen, are given by Whitman (50), who coöperated in designing the test.

(d) *O'Connor Tweezer Dexterity Test*: This test differs from the finger dexterity test in requiring the insertion of only one pin in each hole, tweezers instead of fingers being used. O'Connor (Appendix A) lists a number of factory occupations which may be undertaken by men or women who grade A or B in this test. Wells (48) offers a modification of this test which he believes to be more satisfactory than the original form proposed by O'Connor.

## BATTERIES OF MOTOR TESTS

Single tests of motor ability often gain in usefulness when included in a battery of tests. Brace's team of body-balancing tests has already been discussed on p. 45. Two other test groups, which include tests described above, have been devised: the one by Garfiel (15), the other by Seashore (41). These will be briefly outlined.

**1. Garfiel's Motor Agility Tests**

Garfiel was interested in constructing a scale of motor tests which might serve as a valid index of ability in gymnasium and athletic activities. She investigated sixteen tests, finally choosing eight, which, when combined into a team, gave a high correlation with a criterion of ability in sports. This criterion consisted of an order of merit ranking for athletic ability made by three members of the gymnasium department and by five students. The subjects were fifty sophomores. Six weeks after the first rating, judgments were again obtained from two of the faculty members and from four students. The correlation of .92 between the first and second ratings indicated a high degree of consistency in these judgments. The correlation of the team of tests with the criterion was .77, and its correlation with the Army Alpha Intelligence Examination was — .08. The eight tests chosen were the following:

- (a) 100-yard dash
- (b) Picking up paper
- (c) Strength of back
- (d) Steadiness
- (e) Tricks
- (f) Tapping
- (g) Leg strength
- (h) Hand strength

The Garfiel team of motor-agility tests was investigated in connection with the construction of the Minnesota Mechanical Ability Tests (36). Garfiel's battery was intended for women. The Minnesota tests of motor agility, which were developed for twelve-year-old boys, included only the running, back strength, steadiness and right-hand strength tests. The correlation of the Minnesota Agility Tests with the eight separate Minnesota Mechanical Aptitude Tests ranged from .14 to — .16, averaging — .05. The correlation of the agility battery

with Otis I.Q. was — .38. These extremely low correlations, together with the low correlation of the Garfiel battery and Army Alpha, suggest the existence of motor agility as a trait independent both of verbal intelligence and mechanical skill (36).

## 2. The Stanford Motor Skills Unit

This group of tests was assembled for the purpose of predicting success in complex vocational activities. The advantages of the battery are described by Seashore as follows: (a) The reliabilities of the tests range from .75 to .94, with an average reliability of .84; (b) the intercorrelations between the individual tests are low, ranging from — .03 to .48; (c) a unified scoring system is used; (d) in five of the six tests the scoring is automatic; (e) the apparatus is standard. To these advantages may be added the fact that the tests can be given outside the laboratory. The battery was constituted as follows:

- (a) The Koerth pursuit apparatus, described on p. 51.
- (b) Motor rhythm, or precision in following a regular rhythm pattern on a telegraph key.
- (c) Serial discriminator, or speed of finger movements in discriminative reaction to a visual series.
- (d) Tapping (p. 38).
- (e) Spool packing, involving speed in the coördination of the hands.
- (f) Speed rotor (Miles) requiring speed in turning a small hand drill.

The practical application of this battery to vocational selection and guidance is as yet problematical. Seashore (42) has reported an experiment carried out upon fifty male operatives in a knitting mill, in which scores in the tests were correlated with production in the winding of yarn, a high-speed process. The reliability of the production criterion (*i.e.*, yarn winding) is given as .86. The correlation of the individual tests with the criterion ranged from .07 to .36, while the correlation of the whole battery with the criterion was .14.

Seashore interprets the low intercorrelations of the tests in his battery, as well as their low relationship to his criterion, to mean that motor skills are highly specific. This is an interesting finding from the standpoint of the organization of abilities. Unfortunately, however, it renders doubtful the value of a battery of simple motor tests in forecasting complex manual activities.

## TESTS OF MECHANICAL APTITUDE

Motor ability is usually to be distinguished from mechanical aptitude. Tests of motor ability, as we have seen, are designed to measure functions demanding, in general, speed and precision of movement, hand-eye coördination, and muscular control. Tests of mechanical aptitude, on the other hand, are more complex and much broader in scope. Not only do they draw upon inherited motor abilities, but upon acquired skills and developed interests as well. Expressed more concretely, such tests attempt to measure the ability making for success in tasks requiring the use of tools, the understanding and operation of machinery, skill and dexterity.

The Minnesota investigators (36) set up certain criteria of mechanical ability which will serve to clarify the concept. These criteria were (a) *quality* of mechanical work as judged by experts; (b) *quantity* of mechanical work when a given standard of excellence is maintained; (c) *creativeness* in mechanical work as shown in the devising of new methods or in the utilization of old methods in novel ways; (d) *critical appreciation* of mechanical work which involves, among other things, good judgment in the evaluation of constructed products and a keen interest in mechanical things.

The pioneer attempt to measure mechanical aptitude was the group of tests devised by Stenquist (44), the Assembling Tests and the Mechanical Aptitude Tests. Both of these batteries have found extensive use in schools, as well as in vocational guidance and selection.

### 1. The Stenquist Assembling Tests<sup>1</sup>

This battery embraces three series or groups. The first and second series were intended for children in the upper grades, for high-school pupils, and for adults. The third series was constructed for children in the lower grades.

*Series 1:* This test, which consists of a box divided into ten compartments, is shown in Figure 12. Each compartment contains a single mechanical device, such as a paper clip, a bicycle bell, a clothespin, a mouse trap, and the like. These devices are taken apart, placed in the box, and presented to the subject, who is simply told to put the parts together. The time taken for this series by an average seventh-grade boy is thirty minutes. The reliability of this series ranged from .06 in a group of fifty-two sixth-grade girls to .80 in a group of

<sup>1</sup> May be purchased from the C. H. Stoelting Co., Chicago, Illinois.



thirty high-school boys. Most of the reliability coefficients are above .60.

*Series 2:* This is another series of ten models, to be given in the same way as Series 1. Series 2 correlated .59 with Series 1, in a group of 369 seventh- and eighth-grade boys. In a group of graduate students, men and women, the correlation of the two tests was .75.

*Series 3:* This is a series of ten models, to be used with children in the third, fourth, fifth and sixth grades.

Stenquist (44) has reported a correlation of .40, in a group of 100 seventh- and eighth-grade boys, between Series 1 and a composite intelligence score made up of the Haggerty Intelligence Test,

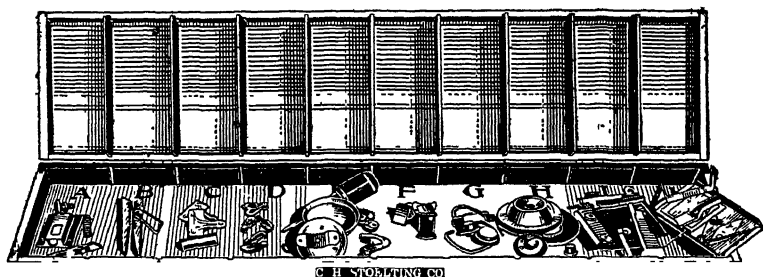


Figure 12.—STENQUIST ASSEMBLING TESTS, SERIES 1  
(Reproduced by courtesy of the C. H. Stoelting Co.)

National Intelligence Tests 1 and 2, Myers Mental Measure, and the Kelley-Trabue Test. The correlation between Army Alpha and Series 2 in the same group was .34. In five army groups, numbering 30 to 216 men, Stenquist obtained correlations between Series 1 and Army Alpha ranging from .00 to .35. Toops (46), who gave the Stenquist Assembling Test to 145 high-school boys, obtained a correlation of .14 with Army Alpha. The correlations between the Stenquist Assembling Test and the Thorndike Intelligence Examination in a group of eighty-two university students were .24 for men, and .13 for women.

The correlations reported above indicate very little relationship between general intelligence and mechanical ability as measured by Stenquist's tests. Because of this low relationship, Stenquist concludes that "an individual's position in General Intelligence is . . . largely independent of his position in General Mechanical Ability and Aptitude." The validity of the Assembling Test is high, as measured by

school marks in shop work. The correlations of Series 1, with teachers' rankings of boys for achievement in shop work, range from .80 to .90 (44).

The Stenquist Assembling Tests have been revised at the University of Minnesota, the revision being known as the Minnesota Assembling Test (36). Many of the mechanical devices used by Stenquist have been retained, and several new ones added. The reliability of the Minnesota Assembling Test was .94 in a group of 150 seventh- and eighth-grade boys. In the same group the correlations between the Minnesota Assembling Test and various criteria of mechanical ability (36) ranged from .24 to .55.

## 2. The Stenquist Mechanical Aptitude Tests I and II<sup>1</sup>

The Stenquist Assembling Test is not adapted to group testing. It is somewhat cumbersome to use and is expensive if many sets are required. To meet these difficulties, Stenquist has devised two paper-and-pencil tests which present a large number of mechanical problems, and involve a minimum of language. In the first of these tests, Test I, ninety-five problems are presented by means of pictures. S is required to indicate which one of five pictures belongs with each of five other pictures. These pictures deal with common mechanical devices, and do not require special training or skill. Test II is much like Test I, but its questions deal more specifically with machinery and machine parts. Familiarity with and interest in machinery are measured more directly by this test. Test I had a reliability of .79 in a group of 103 boys in grades six, seven and eight; while Test II had reliabilities of .61 and .68 in a group of 200 boys in the same grades (44).

The correlations of the Mechanical Aptitude Tests and general intelligence are no higher than those between the Assembling Test and intelligence. Kefauver (26) has reported a correlation of .35 between the Mechanical Aptitude Tests I and II, and the Terman Group Test of Mental Ability, in a group of 101 pupils in Smith-Hughes trade courses. This result agrees with the finding of Paterson (36), who obtained a correlation of .36 between Test II and an average of Army Alpha and Otis I.Q.'s in a group of 217 junior-high-school boys. The studies of Paterson and Kefauver support the conclusion of Stenquist (44) that, generally speaking, intelligence and mechanical ability are largely independent functions.

<sup>1</sup> Sold by the World Book Co., Yonkers, New York.

According to Stenquist (44), the correlation of the Mechanical Aptitude Test I with the Assembling Test, Series 1, is .64, while its correlation with the Assembling Test, Series 2, is .69. The correlations between the Mechanical Aptitude Test II and teachers' rankings of students in shop work range from .43 to .84, averaging about .64. Kefauver (26) found correlations between teachers' ratings of 101 boys for achievement in shop courses, and the Mechanical Aptitude Tests I and II ranging from .07 (automobile mechanics) to .65 (machine shop). When all of the students were taken together without regard to the type of shop work pursued in school, the correlation of test scores with teachers' ratings was .45. These results suggest that Stenquist's tests should be distinctly useful as an aid in the selection of students who possess mechanical aptitude.

### 3. Toops I.E.R.<sup>1</sup> Assembly Test for Girls

Toops (46) has constructed the I.E.R. Assembly Test for girls, with the aim of duplicating with different materials the test situations presented to boys in the Stenquist Assembling Tests. There are eleven tests in the I.E.R. series, typical of which are stringing beads, cross-stitching, trimming paper, inserting tape, *etc.* The correlations between the I.E.R. tests and the Stenquist Mechanical Aptitude Tests I and II in a group of girls, ages twelve to fifteen, inclusive, were as follows: for 76 twelve-year-old girls, .48; for 120 thirteen-year-old girls, .47; for 83 fourteen-year-old girls, .42; and for 39 fifteen-year-old girls, .31. The correlations between these two batteries were somewhat higher for boys; the correlation of the I.E.R. test with the Stenquist Assembling Tests being .53 in a group of thirty seventh-grade boys, and .50 with the Stenquist Mechanical Aptitude Tests I and II combined, in the same group. These results indicate that performance in the I.E.R. Assembly Test depends to a fairly substantial degree upon the same abilities measured by the Stenquist tests.

### 4. Minnesota Mechanical Ability Tests<sup>2</sup>

The most thorough investigation of mechanical abilities has been conducted at the University of Minnesota, and has been described in detail by Paterson, Elliott, *et al.* (36). The possibilities of a large number of tests were investigated, and several batteries of mechanical ability tests were constructed. The criteria against which the

<sup>1</sup>Institute of Educational Research of Teachers College, Columbia University. Sold by the C. H. Stoelting Co., Chicago, Illinois.

<sup>2</sup>Sold by the Marietta Apparatus Co., Marietta, Ohio.

tests were validated consisted of: (a) The finished products made by boys in introductory courses in shop work; (b) the results of tests of final operations in the same courses; (c) the results of objective information tests also given in the same courses. In drawing up a final criterion, account was taken of both quality and quantity. Three measures were obtained: (a) a measure of the *quality* of the work done, known as the "quality criterion"; (b) a measure of the *quantity* of the work done in relation to its quality, known as the "quantity-quality criterion"; and (c) a measure of information about tools and materials and their uses, known as the "information criterion."

A full report of the construction of the Minnesota Mechanical Ability Tests has been given by its authors and need not be repeated here. A few of the batteries, however, will be listed, and some typical results cited.

(a) One of the most extensive batteries consisted of the following tests:

- (1) Minnesota Paper Form Board
- (2) Academic grades
- (3) Minnesota Assembly Test
- (4) Boy's mechanical operations (a questionnaire concerning the boy's mechanical operations in and about the home)
- (5) Interest Analysis Blank (Hubbard revision of Freyd's blank)
- (6) Otis Intelligence Test

The correlation of this team with a combined quality and information criterion was .81, this being the highest validity coefficient between any selection from all of the available mechanical tests and acceptable criteria.

(b) Another battery consisted of the Minnesota Assembly Test, the Interest Analysis Blank, and the Minnesota Spatial Relations Test, which includes four formboards. The correlation of this battery with the quality criterion was .67. In the validation of these batteries the subjects were 100 boys in the seventh and eighth grades, their ages ranging from  $12\frac{1}{2}$  to  $16\frac{1}{2}$ . The correlations of the test scores with chronological age ranged from — .21 to .19.

A number of other batteries were investigated, validity coefficients ranging from .55 to .66. The reliability coefficients of the tests included in these batteries ranged from .86 to .94, in groups of from 100 to 217 junior-high-school boys. On the whole, the Minnesota

Mechanical Ability Tests are satisfactory from the standpoints of both reliability and validity. In fact, this investigation gives an excellent picture of the best available techniques used in present-day test construction.

The authors of the Minnesota Mechanical Ability Tests have studied the effects of environmental influences upon mechanical ability. A thorough investigation of the pupils' environment and previous experience was made. Questionnaires relating to economic and cultural status were employed, and information was obtained covering such things as the mechanical skill shown by fathers and sons in work around the house, fathers' and ancestors' occupations, tools owned by the family, *etc.* When these various measures were correlated against the test scores, the highest coefficient, that between test scores upon an apparatus test battery and the mechanical operations reported by the boy himself, was .39. Other correlations between test scores and environmental influences ranged from — .14 to .35. These results do not indicate a very close relationship between environmental influences of a mechanical sort and the scores achieved by the 100 junior-high-school students tested.

Age differences in mechanical aptitude were investigated by giving the tests to more than two thousand subjects ranging in age from eleven to twenty years. The correlations between age and various combinations of tests ranged from .39 to .52 in groups of 690 to 831 subjects, indicating a moderate relationship between age and mechanical ability as measured by the tests.

In several of the tests, notably the Minnesota Assembly Test, boys and men made significantly higher scores than girls and women. In some of the tests, however, *e.g.*, the Minnesota Spatial Relations, the Card Sorting, and Packing Blocks, sex differences were not significant, or else the girls excelled the boys. These results cast some doubt upon the common opinion that boys and men are natively superior to girls and women in all mechanical operations.

An analysis of the test scores made by students taking different courses revealed that engineering students do not excel liberal arts students in mechanical ability tests. Nor were vocational-school boys superior to academic groups in mechanical ability as measured by the tests. These findings suggest that choice of occupation is not always based upon the possession of specific abilities. Also it appears

that the selection by boys of courses in vocational training is not always due to the fact that they possess superior mechanical aptitude.

### 5. Detroit Mechanical Aptitude Examination for Boys<sup>1</sup>

This is a paper-and-pencil test devised by Baker and Crockett (2). It consists of eight parts, as follows: recognition of tools; a tracing test; estimation of errors; information; sorting of pictures of nuts, bolts, washers and screws; spatial relations; functions of parts of machines; operation of machines. Answers are indicated by filling in numbers and letters, twenty-eight minutes being allowed for the whole test. The test was standardized upon 3,255 boys, and norms are given for ages ranging from 8 to 20. Retests upon 193 pupils after an interval of two weeks gave a reliability coefficient of .76. Test scores were correlated with shop teachers' ratings for fifty boys in a technical high school, resulting in a validity coefficient of .64. The correlations of the test with tests of general intelligence range from .30 to .50, and would undoubtedly be lower if age variability were ruled out, since the norms increase markedly with age.

### 6. MacQuarrie Test for Mechanical Ability<sup>2</sup>

This test, devised by T. W. MacQuarrie (31), is another paper-and-pencil test designed to measure mechanical ability. An effort was made to duplicate on paper a number of psychological tests which have been used at various times to measure motor speed and precision. The test consists of seven parts and includes such performances as tracing, tapping and dotting. The separate sub-tests have reliabilities which range from .72 to .86. The reliability of the whole test was reported by the author to be over .90 in each of three groups of subjects, numbering 35, 80 and 250 cases, respectively. The validity of the MacQuarrie Test was determined by correlating scores made on it with teachers' ratings of mechanical ability and the degree of excellence of mechanical work completed. The resulting correlations ranged from .32 to .81, which is fairly good evidence that the test indicates, at least, the presence of mechanical ability as defined by the criteria.

An attempt was made to discover whether the ability measured by the test was related to the ability measured by general intelligence tests. The correlation of the MacQuarrie Test with group intelligence tests was never above .20 and in one group of sixty cases the correla-

<sup>1</sup> Sold by the Public School Publishing Co., Bloomington, Illinois.

<sup>2</sup> Sold by the C. H. Stoeltng Co., Chicago, Illinois

tion was .00. It appears, therefore, that this test measures something very different from the traits measured by group mental tests. The MacQuarrie Test has simplicity of construction and ease of administration in its favor.

## BIBLIOGRAPHY

1. ABELSON, A. R., "The Measurement of Mental Ability of 'Backward Children,'" *British Journal Psychology*, 4:268-314, 1911.
2. BAKER, H. J., and CROCKETT, A. C., *Detroit Mechanical Aptitudes Examinations for Boys and Guls*, Public School Publishing Company, Bloomington, Illinois, 1929.
3. BALDWIN, B. T., AND STECHER, L. I., *The Psychology of the Pre-school Child*, D. Appleton & Company, New York, 1925.
4. BICKERSTETH, M. E., "Application of Mental Tests to Children of Various Ages," *British Journal Psychology*, 9:23-73, 1917.
5. BLACKBURN, J. M., "Individual Differences in the Performance of a Simple Test," *British Journal Psychology*, 21:383-393, 1930.
6. BRACE, D. K., *Measuring Motor Ability*, A. S. Barnes & Company, New York, 1927.
7. BRONNER, A., HEALY, W., LOWE, G., AND SHIMBERG, M., *A Manual of Individual Mental Tests and Testing*, Little, Brown & Company, Boston, 1927.
8. BURT, CYRIL, "Experimental Tests of General Intelligence," *British Journal Psychology*, 3:94-177, 1909.
9. CAROTHERS, F. E., "Psychological Examinations of College Students," *Archives Psychology*, No. 46, 1921.
10. CARVER, D. J., "The Immediate Psychological Effects of Tobacco Smoking," *Journal Comparative Psychology*, 2:279-302, 1922.
11. COWDERY, K. M., "A Note on the Measurement of Motor Ability," *Journal Educational Psychology*, 15:513-519, 1924.
12. DEWEY, E., CHILD, E., AND RUMML, B., *Methods and Results of Testing School Children*, E. P. Dutton & Company, New York, 1920.
13. DUNLAP, K., "Improved Forms of Steadiness Tester and Tapping Plate," *Journal Experimental Psychology*, 4:430-433, 1921.
14. FRANZ, S. I., *Handbook of Mental Examination Methods*, The Macmillan Company, New York, 1919.
15. GARFIEL, E., "The Measurement of Motor Ability," *Archives Psychology*, No. 52, 1923.
16. GATES, G. S., "Individual Differences as Affected by Practice," *Archives Psychology*, No. 58, 1922.
17. GOODENOUGH, F. L., AND BRIAN, C. R., "Certain Factors Underlying the Acquisition of Motor Skill by Pre-school Children," *Journal Experimental Psychology*, 12:127-155, 1929.
18. GRIFFITTS, C. H., "A Study of Some Motor Ability Tests," *Journal Applied Psychology*, 15:109-125, 1931.

19. HERTZBERG, O. E., "Relationship of Motor Ability to the Intelligence of Kindergarten Children," *Journal Educational Psychology*, 20:507-519, 1929.
20. HOLLINGWORTH, H. L., "The Influence of Caffeine on Mental and Motor Efficiency," *Archives Psychology*, No. 22, 1912.
21. HOLLINGWORTH, H. L., "Correlation of Abilities as Affected by Practice," *Journal Educational Psychology*, 4:405-414, 1913.
22. HOLLINGWORTH, H. L., "The Influence of Alcohol," *Journal Abnormal and Social Psychology*, 18:204-237, 1923.
23. HULL, C. L., "Influence of Tobacco Smoking on Mental and Motor Efficiency," *Psychological Monographs*, No. 3, 33, 1924.
24. JOHNSON, B. J., "Practice Effects in a Target Test—a Comparative Study of Groups Varying in Intelligence," *Psychological Review*, 26: 300-316, 1919.
25. JOHNSON, B. J., *Mental Growth of Children in Relation to the Rate of Growth in Bodily Development*, E. P. Dutton & Company, New York, 1925.
26. KEFAUVER, G. N., "Relationship of the Intelligence Quotient and Scores in Mechanical Tests with Success in Industrial Subjects," *Vocational Guidance Magazine*, 7:198-203, 1929.
27. KITSON, H. D., "Determination of Vocational Aptitudes: Does the Tapping Test Measure Aptitude as Typist or Pianist?," *Personnel Journal*, 6:192-198, 1927.
28. KOERTH, W., "A Pursuit Apparatus: Eye-hand Coördination," *University of Iowa Studies in Psychology*, No. 8, *Psychological Monographs*, 31:288-292, 1922.
29. LANIER, L. H., "Prediction of the Reliability of Mental Tests and Tests of Special Ability," *Journal Experimental Psychology*, 10:69-113, 1927.
30. LINK, H. C., "An Experiment in Employment Psychology," *Psychological Review*, 25:116-127, 1918.
31. MACQUARRIE, T. W., "A Mechanical Ability Test," *Journal Personnel Research*, 5:329-337, 1927.
32. MILES, W. R., "The Pursuimeter," *Journal Experimental Psychology*, 4:77-105, 1921.
33. MILES, W. R., "Static Equilibrium as a Useful Test of Motor Efficiency," *Journal Industrial Hygiene*, 3:316, 1922.
34. MUSCIO, B., "Motor Capacity with Special Reference to Vocational Guidance," *British Journal Psychology*, 13:157-184, 1922-1923.
35. O'CONNOR, J., *Born That Way*, The Williams & Wilkins Company, Baltimore, 1928.
36. PATERSON, D. G., ELLIOTT, R. M., *et al.*, *Minnesota Mechanical Ability Tests*, University of Minnesota Press, Minneapolis, 1930.
37. REAM, M. J., "The Tapping Test: a Measure of Motility," *Psychological Monographs*, No. 1, 31:293-319, 1922.



38. RENSHAW, S., "An Experimental Test of the Serial Character of a Case of Pursuit Learning," *Journal General Psychology*, 1:520-533, 1928.
39. RENSHAW, S., AND WEISS, A. P., "Apparatus for Measuring Changes in Bodily Posture," *American Journal Psychology*, 37:261, 1926.
40. RENSHAW, S., AND POSTLE, D. K., "Pursuit Learning under Three Types of Instruction," *Journal General Psychology*, 1:360-367, 1928.
41. SEASHORE, R. H., "Stanford Motor Skills Unit," *Psychological Monographs*, No. 2, 39:51-66, 1928.
42. SEASHORE, S., "The Aptitude Hypothesis in Motor Skills," *Journal Experimental Psychology*, 14:555-561, 1931.
43. SOMMERVILLE, R., "Physical, Motor and Sensory Traits," *Archives Psychology*, No. 75, 1924.
44. STENQUIST, J. L., *Measurements of Mechanical Ability*, Teachers College, Columbia University, Contributions to Education, No. 130, 1923.
45. TITCHENER, E. B., *Experimental Psychology: Students' Manual, Qualitative*, The Macmillan Company, New York, 1906.
46. TOOPS, H. A., *Tests for Vocational Guidance of Children Thirteen to Sixteen*, Teachers College, Columbia University, Contributions to Education, No. 136, 1923.
47. WELLMAN, B., "The Development of Motor Coördination in Young Children," *University of Iowa Studies in Child Welfare*, No. 4, 3, 1926.
48. WELLS, F. L., "Comparative Reliability in Tests of a Motor Aptitude," *Journal Genetic Psychology*, 37:318-320, 1930.
49. WHIPPLE, G. M., *Manual of Mental and Physical Tests. Simpler Processes*, Warwick and York, Baltimore, 1914.
50. WHITMAN, E. C., "A Brief Test Series for Manual Dexterity," *Journal Educational Psychology*, 16:118-123, 1925.

## CHAPTER III

### TESTS OF PERCEPTION AND ATTENTION

IN EVERY-DAY life, those objects and events perceived or experienced by way of the senses furnish the primary data which are the basis of mental activity. In a real sense, therefore, the function of perception may be thought of as fundamental to all behavior. To perceive is to grasp, *i.e.*, to apprehend and to discriminate. Hence tests of perception are concerned primarily with the individual's ability to perceive likeness and difference; with the amount of material which he can apprehend, and the speed with which he reacts; and with relationships of space, time, form, color, *etc.*

Strictly speaking, there are no tests of attention *per se*, but achievement in any mental test depends in part upon attention. This is especially true of perception tests, in which a successful response demands concentration of a high order. Tests of perception, therefore, are at the same time particularly good measures of attention, and have often been employed for this purpose. The following sections will describe some typical tests of perception and attention.

#### 1. Span of Visual Perception

The greatest amount of material that can be visually apprehended in a single moment of attention is called the span of perception. The typical procedure in measuring span of visual perception is to expose cards containing the test material for a period of .1 second ( $100 \sigma$ )<sup>1</sup> or less. This short exposure interval, which is necessary in order to preclude eye movements, demands the use of a tachistoscope or other automatic exposure device. (See Figure 13.) In a tachistoscope each card is shown for an instant only, through an aperture, and is immediately covered by a shutter. S is required to fixate a point upon the concealing shutter. When his attention is fixed he signals to the experimenter, who exposes a card. Another method is to give S a warning signal, and after a prearranged interval, say two seconds, to expose

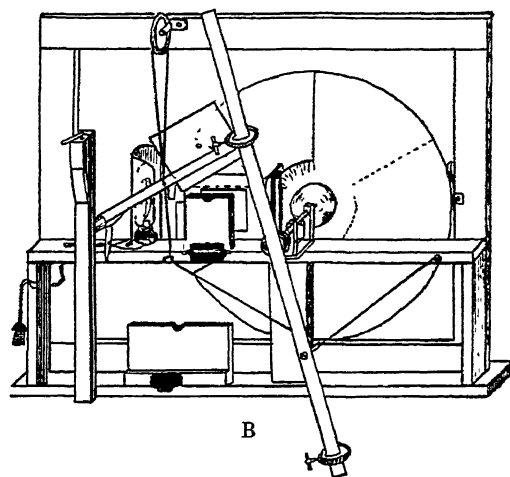
<sup>1</sup> $\sigma = .001$  second.

a card. This preparatory interval is intended to establish a more definite mental set than would be possible with an indefinite interval.

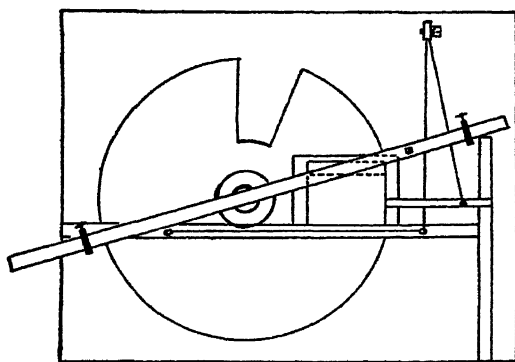
The earliest experiment upon visual perception span was performed by Cattell (9) who reported the normal span for adults to be four digits or three to four letters. Whipple (55) gives the average adult span as 4.8 letters. He states that ordinarily four to five unrelated objects can be grasped in a single exposure, though many individuals do better than this. Much depends upon the kind of material employed, and upon the age and training of the subjects. For related material the span of visual perception is longer than for unrelated. Many individuals will grasp fifteen to twenty letters when they form familiar words, and a smaller number when the letters are grouped into unfamiliar words. As many as thirty-two letters can be read when they make up a sentence of short words. These facts are important in the psychology of reading (p. 69).

According to O'Brien (33), the span of perception in silent reading increases with age and with school grade. From his curves of growth in perception span ([33], p. 99), it appears that improvement is rapid up to the end of the fourth grade, the average number of eye fixations per printed line decreasing from 18.6 (Grade I) to 7.5 (Grade IV). From grade four on to the first year of high school, there is a plateau, which is followed by a second rise, the average number of fixations per line for the college group being 5.9. Similar results have been recorded by Gray (23) who found an increase in perception span beginning with the third grade, erratic results in the fourth and fifth grades, and from then on a slow increase to the college level.

Except at the very beginning, the span of perception in the case of adults increases little, if at all, with practice. One of the earliest experiments upon the effects of practice was performed by Whipple (54) with six college students as subjects. Test material consisted of groups of unrelated letters, five, six or seven in number which were exposed for .08 second. The number of letters perceived at the beginning of the experiment lay between four and five. After a practice period of from seven to ten days, the perception span showed a slight increase, which Whipple attributes to the better adaptation of his subjects. Relatively slight improvement was found also when dots, pictures, drawings, nonsense syllables and stanzas of poetry were used as exposure material. Foster (17) has performed a some-



A



C

Figure 13.—WHIPPLE DISC TACHISTOSCOPE

A, front view; B, rear view; C, front view, screen removed.

what similar experiment to that of Whipple's, using objects, pictures and nonsense drawings as material. His subjects were three trained psychologists, and the exposure time varied from ten to sixty seconds. When exposure times are as long as this, so that eye movements are allowed, the response is called the *span of apprehension* rather than the span of perception. Foster found that the increase in ability to reproduce material appeared early in the experiment, the limit being reached in a short time. He attributes the improvement found to the acquisition of more efficient work habits, such as tricks of counting, grouping, *etc.* There was little evidence that practice actually increased the span of apprehension.

Dallenbach (13) has investigated the perception span of twenty-nine second-grade children of both sexes with somewhat different results from those cited above for adults. Ten minutes' drill was given daily for seventeen weeks, digits, letters, combinations of both, geometrical forms, *etc.*, being used as test material. Dallenbach's results may be summarized as follows: At first there was rapid improvement followed by a marked loss in rate of improvement; the increase in perception span, however, persisted even after forty-one weeks of no practice. Since perception span in reading is known to increase during the early ages (33) this result may be attributed in part to the fact that the children in this experiment were nearly one year older when finally tested.

The relation of span of visual perception to intelligence has received some attention from investigators. Dallenbach (14), in a study of eighteen girls and twenty-three boys, all of whom were inmates in an institution for the feeble-minded, obtained a correlation of .70 between visual apprehension and mental age found by the Yerkes-Bridges Point Scale. The wide age range of Dallenbach's subjects (ten to eighteen years) makes it probable that this correlation would be lower if age variability were held constant. Tinker (49), who has made an exhaustive summary of the work done upon visual apprehension and perception in reading, quotes Ranschburg (42) to the effect that feeble-minded children have a definitely smaller perception span than have normal children. In another experiment cited by Tinker, Hoffman (27) found a correlation of .60 between general intelligence and perception span in a group of 250 grade school children. These studies indicate a substantial relationship between general mental ability and span of perception in

the early years. This is to be expected, perhaps, since in children perception span is a measure of alertness and of interest as well as of "intake."

In the psychology of reading the span of visual perception is of fundamental importance, although a wide visual span does not guarantee efficiency in reading. Gray (24) has shown that it is possible to have a wide span of perception and yet be an inefficient reader. Buswell (6) who worked with 186 students in grade school, high school and college, reported a positive correlation between perception span and "mature reading habits." Mature reading habits, as defined by Buswell, embrace the following characteristics: Speed of perception, few fixations per line, few or no refixating movements and minimal duration of fixation. It is well known (Dearborn [15]) that in reading, the eyes do not move smoothly along the line, but proceed in jerks from left to right, with short fixation pauses between movements. It is during these pauses that the reading matter is perceived; and hence a wide visual span permits the apprehension of more material than does a narrow one. In reading it is important to reduce not only the number of fixation pauses per line, but the duration of the rest pauses as well. Efficiency in reading demands, in addition, a minimum of backward movements along the line. A discussion of methods of studying these factors and of results obtained is given by Gray (24) and by O'Brien (33).

#### Test 1. Span of Visual Perception

*Apparatus:* Whipple disc tachistoscope, or any standard short exposure apparatus; set of cards containing varying numbers (3-10) of dots, digits, nonsense syllables, unrelated words, easy words, familiar words, etc. The exposure material may be made by the experimenter, or purchased from the C. H. Stoelting Company, Chicago, Illinois. The disc tachistoscope is obtainable from the C. H. Stoelting Company, Chicago, Illinois.

*Method:* First adjust the aperture of the tachistoscope for an exposure of .1 second. Whipple (55) gives the method for adjusting the disc tachistoscope for exposure of different lengths. If this instrument is used, have S rest his chin upon the chin rest and focus his eyes upon the point in the center of the fixation card. Insert the first card in the holder and when S's full attention is secured, release the trigger and expose the card. Write down S's report. The trigger may be released upon signal from S, or after a predetermined interval, say, two seconds after the "Ready" signal. By attaching a string to the trigger, it is possible for S to control the interval before exposure himself.

*Record:* Most investigators take the span of visual perception to be simply the largest number of items that can be grasped by S, without error, in a single exposure. Three trials may be given on each card until S fails in all three; or the same card may be presented over and over again, the number of trials needed for a correct response being recorded. *Norms:* Norms determined by different investigators will be found on p. 66. In the Columbia University Laboratory the following result has been secured: With fifty-three men and women subjects, using the disc tachistoscope, with an exposure time of .1 second, the perception span for unrelated letters averaged 5.09, with an S.D. of 1.58.

## 2. Cancellation

Tests of cancellation have been used to measure perception, attention and even speed of movement. There are many varieties of the cancellation test, but all are alike in requiring the subject to mark out certain recurring (and designated) items from a page of material, with maximum speed and accuracy. The material, for example, may consist of a sheet containing letters printed in random sequence, through which S is instructed to go, line by line, and cross out every *a* that he finds. The task may require the cancellation of two or more letters, say, every *a* and *t*, or *a*, *t*, *c* and *g*; or digits may be used instead of letters, the instructions being to cancel single digits or combinations of digits. Again S's task may be to cross out every four-letter word on the page. In such tests the words may be spaced as in ordinary reading matter, or printed with no spacing between. Still other forms of this test call for the crossing out of circles, squares, crosses, *etc.*, from sheets containing a variety of such forms.

Few tests have been employed for so many purposes as have cancellation tests. Scores in cancellation have been correlated with measures of general intelligence, with vocational skills and with special abilities of various sorts. Cancellation has been used as a measure of the effects of various kinds of motivation; of the effects of smoking, of different degrees of attention and of various mental sets. Cancellation tests have been employed also in studies of sex differences, in the comparison of twins, and in the estimation of time intervals.

The correlations between tests of general intelligence and cancellation are rarely higher than .30 for children, and in the case of adults are often lower (.55). Hertzberg (26) gave two cancellation tests to forty-six kindergarten children. The first test consisted simply of crossing out circles in order upon a checkerboard of circles, the

score being the number of figures crossed out in one minute. The correlation of the scores in this test with Stanford-Binet Mental Age was .55. This coefficient dropped to .29, however, when the variability introduced by differences in age was partialled out. The second test consisted in crossing out circles on a sheet containing stars, circles, squares and triangles arranged in random order. These cancellation scores correlated only .19 with mental age when variability arising from differences in chronological age was held constant. Brown (5) compared the cancellation scores of thirty-nine girls and forty boys, eleven to twelve years of age, with teachers' estimates of intelligence and with school grades, obtaining correlations ranging from .00 to .28. Vickery (51), in a recent experiment, gave the *A* cancellation test to eighty women in Alabama College. The work period was twenty minutes, a longer time than is usually devoted to cancellation tests. The correlation between the Otis Self-Administering Test of Mental Ability and cancellation in this group was .25. Sommerville (46) who correlated the cancellation scores of 102 college men with their scores in the Thorndike Intelligence Examination obtained a coefficient of .16.

The cancellation test exhibits a fair degree of correlation with those mental tests which demand quickness and discrimination. Garrett and Lemmon (18) have reported correlations of about .40 between cancellation scores and tests of analogies, completion, and word building in adult groups. From tests given to six classes in grades four and five in the Horace Mann School, Gates (20) obtained the following correlations:

Cancellation with:	<i>r</i>
Comprehension of reading	.28
Thorndike-McCall Reading Test	.27
Courtis Rate of Silent Reading	.30
Courtis Comprehension of Reading	.29
Monroe Comprehension of Reading	.16
Word Perception	.26
Picture Naming	-.03

Gates's cancellation test consisted of rows of numbers (mostly five digits each) paired against each other. The task was to cross out each number which was not identical with its paired number. The correlations of the reading tests with cancellation are fairly consistent but not very high. The Picture Naming Test, which is quite different from cancellation, has a low negative correlation with it.



## Cancellation Test

hplgvjembsfgtcdbvmzkhfpoiabgjflurcqihdjoabkvt  
 ndefxkjcdtmwfzeojqlfhycijwpzhkeqfvyzlsxfpvrjy  
 mxniufktvxpyralkjowqfvpystexralpbicqrdjfuqzihg  
 pskdcmosgfyqwepkasditogmqkftshbdrpzvxqufsid  
 tohxwaklbvxzfoearlsjvqfuoltdapqevkmtpuodszejwg  
 xfvozpqrbeftkxrvjybuacdsbwumehrcdxygjwhblft  
 yekdwzvxbpokwizyedgowacpkmjrhltoidaxkhmwz  
 ytejscqioxhtfayubltrezipwmslbgvexniwoybhzfkm  
 tndrbuclmteazyjgivptwohswfzyqrhlnajyozptqkba  
 msocfvukbijpcyfaoujzmcikrvxptcndsoaljyrbcvzsgu  
 yknbfcgzjprinqkdfgawulcrkgdfuiqkaczymdlxfqokz  
 nwujgredlzpnpjxgdfzemojndxcuizwbjqdghvusi qod  
 zbuecgtpqkuwljorbkspwujtoebnmwadfsvnkubroljw  
 icuofnaedjxcgvznlfsjbhzruxnfmodyhcvsqrukfgdaie  
 xmphfdoqcretbmivlcdnfhqkgpxoasvyn txwbqpfhvn  
 ekdzjrylqamspkkgzonieusyrpnqzcsvtuiygfkpzsnldt  
 vkbpasynjxhobuywqzcljgyrovdwnmuqkfxihyvga j  
 wlusnmxbgtyikpaugxltymbingujxsrpnmhgvbzluxe  
 wgtjfablhxmqsynaziegmsyqjdnxlebztagmahfioqus  
 rtmhlciqagnyrtcljhinrbmqepglcsvmrwxpcqzbtaks n  
 yvgxzodcabeximnfyutjrxnzhwb dipcolevxintmh wpl  
 iymgdwrakivqyselwdfikuhvsgdcimwopvqkinudwr  
 htlvxsizclqwpjuzraehwivsrmyubevpnigoqsmvzer  
 uysmxanigzveqxrpwosnltxzujwgakmehlcy nbkpww  
 oahyispktgcmnqzsejiucodrh esbgufmxopnqckegmu  
 xatkojhbeavupsrmtxqwk hbcoryuaqhimzkwbnqjria  
 ldevcbtqnwlxdrz fecgrpinajhxtqkyiwmucgolkebam  
 qrgnvjswhdexcormuplhq rnzetwblhcg rtjamlfhycxr  
 bovdznyheilavcfonyh xatzwgnhjfmowbpxhtsd fvep  
 wimngsaectjqwhftypdos.

Name.....

Date..... Sex.....

Age..... Grade.....

Test 26

Speed in cancellation increases with practice, according to the results of a number of investigators. DeWeerd (12) found that forty-five children, practicing three minutes daily for eleven days on the Woodworth-Wells Cancellation Test, improved as much as 62 per cent. in speed. Gates (21) gave the Woodworth-Wells Number Cancellation Test to about 200 fifth- and sixth-grade children. If we take the record of the first day as a base, the gains for the four days following were 12 per cent., 19 per cent., 26 per cent., and 33 per cent. The cancellation test lends itself readily to studies of rates of work and limits of improvement, and has been so used by a number of workers. It is also useful in the investigation of individual differences in learning (Kincaid [30]).

Several attempts have been made to find a relationship between cancellation and vocational skills. Link (31), in a study to which reference has already been made (p. 39), reported that inspection of shells had a correlation of .63 with the Woodworth-Wells Number Cancellation Test; while gauging shells yields a correlation with cancellation of only .17. The subjects in this experiment were fifty-two inspectors and twenty-one gaugers (all girls) employed in a munitions and arms factory. Evidently the cancellation test made a real distinction between these two occupations.

The reliability of the cancellation test is usually satisfactory. The Minnesota investigators (36) allowed their subjects two minutes for number cancellation, and two minutes for letter cancellation. In a group of 217 boys in the seventh and eighth grades, the reliability of number cancellation was .76, and letter cancellation .80. The tests were the Woodworth-Wells Number Cancellation blanks (57). Carothers (8), who gave the Woodworth-Wells Number Cancellation blank to a group of forty-five freshmen girls, obtained a reliability of .60. A summary of early studies of cancellation has been given by Whipple (55), who reports the reliabilities of the test to range from about .60 to .97, for various forms and lengths of the test. The highest reliabilities are those reported by McCall (32), who gave the Woodworth-Wells Cancellation Tests to a group of eighty-eight sixth-grade children. Cancellation of the letter *S* for four minutes gave a reliability of .93, while cancellation of the letter *A* for seven minutes gave a reliability of .95. Two number cancellation tests had reliabilities of .96 and .97, when eight minutes were allowed for

each test. The reliability of a cancellation test evidently depends directly upon its length, or upon the time spent upon it.

## Test 2. Cancellation

*Materials:* Printed forms of cancellation tests of various sorts, *e.g.*, digits, letters, geometrical figures, *etc.*; a stop-watch; a large time clock (if the work-limit method is to be used). A large variety of cancellation blanks may be obtained from the C. H. Stoelting Company, as well as a stop-watch and time clock. A letter cancellation blank is shown in Figure 14.

*Method:* Cancellation tests may be conducted either by the time-limit method or by the work-limit method. In the first method, two minutes or less are allowed and a record taken of the number of items cancelled. In the second method, S is allowed to work through the entire sheet, and the time taken to complete is recorded. Unless the large time clock, from which each subject reads his own score, is employed, it is difficult to work with more than one subject at a time, if the work-limit method is used.

To administer the test by the time-limit method, pass out the test forms, say the letter cancellation test, face down. Instruct the subjects to turn their blanks over at the *go* signal and to cross out as rapidly as possible every letter *a* on the page. At the expiration of two minutes, say *stop*. In the work-limit method each subject, as soon as he has finished the blank, looks up immediately and reads his time from the clock, which was started at the signal *go*.

*Record:* For most purposes, it is sufficient to take as score the total number of items crossed out in the time allowed. Errors are infrequent, and if they occur may be subtracted from S's score. If the work-limit method is used, the score is the number of seconds taken to complete, plus a penalty of one or more seconds for each item omitted.

*Norms:* Norms for adults on several cancellation tests are given by Whipple (55). The following results have been secured in the Columbia Laboratory from groups containing both men and women:

Test	N	
A . . . . .	28	Average time to complete: 82.90 seconds S.D. 15.21
a-t . . . . .	90	Average number of cancellations in 120 seconds: 32.49 S.D. 6.25
Number Group Checking	49	Average time to complete: 151.86 seconds S.D. 18.85

Norms for children are published by Pyle (39). Table XII, which contains useful norms for children, is taken from Dewey, Child and Ruml (16).

TABLE XII

## CANCELLATION TIME (CHILDREN)

Time (in seconds) required to cancel all *a*'s in *hpl* and *xy* blanks

(From Dewey, Child and Ruml [16])

## Boys

Age	9 0-9 9	10 0-10 9	11 0-11 9	12 0-12 9	13 0-13 9
Mean	327 ± 12	303 7 ± 6 7	281 1 ± 6 8	261 2 ± 7 8	247 7 ± 5 9
S.D.	120 0 ± 8 1	69 2 ± 4 7	69 0 ± 4 8	80 4 ± 5 3	61 1 ± 4 2

## Girls

Age	9 0-9 9	10 0-10 9	11 0-11 9	12 0-12 9	13 0-13 9
Mean	319 4 ± 8 1	285 ± 11	287 7 ± 7 3	257 9 ± 6 9	227 4 ± 6 1
S.D.	81 9 ± 5 9	113 0 ± 7 9	73 7 ± 5 2	72 0 ± 4 9	63 8 ± 4 3

## 3. Card Sorting

Card sorting has often been used to measure the influence of various factors upon motor learning and the effects of practice. In this section, however, we are chiefly interested in this test as a measure of sensory-motor discrimination and of visual perception. Card sorting as a test of perception is somewhat akin to the cancellation test, in that both demand quick movement and rapid judgment of differences.

The influence of age and other factors upon card sorting has been investigated in several studies. In a study by Wellman (53) which included fifty-four children three to six years of age, a correlation of .65 was obtained between card sorting and the tracing board. When variability arising from differences in chronological age was partialled out, this correlation dropped to .32. When variability due to mental age (Stanford-Binet) was held constant, the correlation between tracing and card sorting fell from .65 to .08. These results show clearly the influence of physical and mental maturity upon performance in card sorting. In another series of experiments, Wellman found a correlation of .86 between card sorting and the tracing-path performance (p. 50). The experimental group in this study consisted of 136 children. When variability arising from differences in mental age was held constant this correlation fell to .49.

The subjects in all of these experiments were children. A study of card sorting in which adults were subjects has been reported by

Calfee (7). The subjects were 103 freshmen in the University of Texas (fifty-one men, fifty-two women), and thirty boys in the grade schools. The task was to sort fifty cards into five piles, in accordance with the numbers printed on the cards. Correlations were calculated between the card-sorting test and each of five other tests, *viz.*, card dealing, alphabet sorting, a mirror test, a test of vital capacity (p. 12) and school grades. Coefficients of correlation ranged from —.16 for card sorting and vital capacity, to .71 for card sorting and card dealing. The correlations of card sorting with school grades were .50 for the grade-school boys, .29 for the college men, and .28 for the college women. It seems clear that card sorting is a test of learning for children to a greater extent than it is for adults.

The correlation between card sorting and tests of general intelligence is slight, but is higher for children than for adults. Atkinson (2) administered the card-sorting test to thirty college students, of whom eighteen were men and twelve women. The sorting box contained thirty pigeonholes numbered to agree with the numbers on the cards. There were 150 cards in the deck, and seventy-five trials were given, distributed over twenty-five practice periods. Correlations were calculated between each of the practice periods and scores on the Army Alpha Test. With the exception of the first practice period, which gave a correlation of .39, the coefficients ranged from —.10 to .11. Garrison (19), who gave the card-sorting test along with the Otis Self-Administering Test of Mental Ability to sixty college students, obtained a correlation of —.02. In the Minnesota investigation (36) the correlation between card sorting and an average of Otis I.Q. and Army Alpha I.Q. was .15, the subjects being 217 boys. Ruch (44), who worked with grade-school children, has obtained results that are quite similar to these. Ruch instructed fifty-two pupils in the seventh, eighth and ninth grades to sort 100 cards bearing unfamiliar designs into ten compartments. Five trials daily were given for nine days, and the results correlated against Stanford-Binet mental age. For the first trial only the correlation was .33, when variability arising from chronological age differences was held constant. The correlation of the average of the first five trials, *i.e.*, the first day's work, with M.A. was .18. The correlations of card sorting and mental age for the succeeding nine days ranged from —.02 to —.16. Except for the first trials, which apparently require a certain degree of intelligent adaptation, it appears fairly

certain that, even for children, the card-sorting test is not significantly related to measures of abstract intelligence.

Several studies have been made of the relationship between card sorting on the one hand and vocational skill and special aptitude on the other. Link (31), in his study of the inspecting and gauging of shells (see p. 39), used a card-sorting test. The pack to be sorted contained forty-nine cards, each card bearing from seven to twelve letters. Twenty of the forty-nine cards contained the letter *O*. S was asked to sort the cards into two piles, putting into one pile cards which contained the letter *O*, and into the other those which did not. Scoring was in terms of time and errors. The correlation between vocational efficiency and card sorting was .55, for fifty-two inspec-

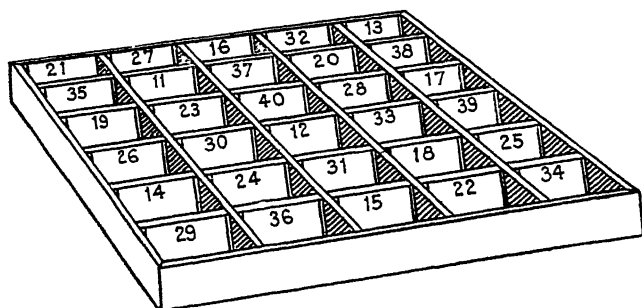


Figure 15—CARD SORTING BOX

tors; and .05 for twenty-one gaugers. This is an interesting finding. It appears that the functions involved in card sorting are fairly closely related to the work of inspecting shells, which demands quickness of movement and hand-eye coördination; but not to gauging in which precision, rather than speed and coördination, is stressed.

The card-sorting test has been used for many purposes. The test has been employed by Baldwin and Stecher (3) in studying the effects of practice; and by Peterson and Barlow (37), by Ruch (43), by Gates (21), by Atkinson (2), and by Crafts (11). These studies have dealt with the effects of practice upon individual differences (Peterson and Barlow); with the correlation of initial and final capacities (Ruch); with speed of response at the various stages of learning (Atkinson); and with the learning process involved in card sorting (Gates and Crafts). Pyle and Snadden (41) have used the

card-sorting test in studying the learning processes of bright and dull children. Obviously, the test lends itself readily to a wide variety of uses. It is possible to use many forms of sorting. Jones (29), for instance, used colored pegs which were to be sorted into glasses. Marbles, stones, sticks, tags and many other objects have been so employed.

The reliability of the card-sorting test will depend directly upon the length and difficulty of the task. In a recent study by Tinker, Imm and Swanson (50) sixty cards, fifteen in each suit, were to be sorted into four boxes. For a group of forty-five university students of both sexes, the reliabilities of the sorting test from trial to trial and from day to day ranged from .81 to .96. When corrected for attenuation, these reliabilities ranged from .90 to .98. In the Minnesota study (36) in which a simpler test was employed, namely, sorting red and blue cards into two compartments, a lower reliability of .72 was obtained.

### Test 3. Card Sorting

*Apparatus:* Sorting box containing compartments suitably marked to agree with the markings upon the cards to be sorted; set of cards. (See Figure 15.) A number of sorting boxes are made by the C. H. Stoelting Company, Chicago, Illinois.

*Method:* Give S the pack of cards, and explain that each card has a marking which corresponds to a marking upon one compartment of the sorting box. Instruct S to begin when the *go* signal is given and to place each card in its proper compartment as rapidly as possible. Allow a time limit, or stop S after a given interval has passed.

*Record:* The score is the number of seconds taken to complete, if the work-limit method is used; or number of cards sorted, if the time-limit method is used. Errors, consisting of placing cards in wrong compartments, may also be reported and allowance made in the score.

*Norms:* The norms given in Tables XIII and XIV are from Pyle (40) and from Dewey, Child and Ruml (16). Other results secured with card-sorting tests may be found in Woolley (58) and Cornell (10).

### 4. The Wallin Peg Boards

These simple tests devised by Wallin (52) are well adapted for use with pre-school children, to whom, according to Stutsman (48), they are exceedingly interesting. The Peg Board Tests consist of four boards, each of which contains six holes into which pegs are to be fitted by the child. Board A is for round pegs only, Board B for square pegs, Board C for three round and three square pegs, and

TABLE XIII  
CARD-SORTING TIME (ADULTS)

In column 1 are the averages of actual scores in seconds for each sorting. In column 2 are the reciprocals of the actual scores. In column 3 is shown the number of cards sorted per minute. The number of subjects was twenty-four, the number of pigeonholes, twenty; the number of cards 100. (From Pyle [40])

	Trials	(1)	(2)	(3)		Trials	(1)	(2)	(3)
1	.....	298	336	20 1	11	. . .	129	775	46 5
2	....	204	490	29 4	12	... .	116	862	51 7
3	...	196	510	30 6	13	. . .	113	885	53 1
4	....	162	617	37 0	14	. . .	114	877	52 6
5	.....	152	658	39 5	15	. . .	111	901	54 1
6	.....	154	649	39 0	16	. . .	111	877	52 6
7	.....	143	699	42 0	17	. . .	105	952	57 1
8	. . .	138	725	43 5	18	. . .	104	962	57 7
9	.. . .	126	794	47 6	19	. . .	102	980	58 8
10	. . .	122	820	49 2	20	. . .	98	1020	61 2

TABLE XIV  
CARD-SORTING TIME (CHILDREN)

Average Time in Seconds Required to Sort Forty-eight Cards into Four Compartments  
(From Dewey, Child and Ruml [16])

Boys

Age	..	9 0-9 9	10 0-10 9	11 0-11 9	12 0-12 9	13 0-13 9
Mean.	..	61 8 ± 1 1	55 27 ± 94	49 94 ± 79	47 63 ± 84	41 91 ± 68
S.D.	..	10 90 ± 75	9 77 ± 66	8 00 ± 56	8 65 ± 60	7 06 ± 48

Girls

Age	.....	9 0-9 9	10 0-10 9	11 0-11 9	12 0-12 9	13 0-13 9
Mean	...	54 44 ± 91	51 9 ± 1 3	51 2 ± 1 1	46 80 ± 85	42 60 ± 63
S.D.	.....	9 18 ± 64	13 00 ± .91	11 10 ± 78	8 85 ± 60	6 62 ± .44

Board D for two round, two square, and two triangular pegs. All of the boards require simple motor coördination, while the last two involve, in addition, simple form discrimination. Boards A and B form part of the Merrill-Palmer Scale of Tests for young children (48). Baldwin and Stecher (3) include all four boards in their series of tests for motor control and form perception.

A comprehensive study of the Peg Board Tests has been made by Goodenough (22), who worked with 300 children, two, three and four years old. There were 100 children (fifty boys and fifty girls) at each age level. The children were given the four Peg Boards and the Kuhlmann Revision of the Binet Tests. The reliability of the



peg boards, as determined by retests after six weeks, varied with the age of the group, being .79 for age 2, .69 for age 3, and .58 for age 4. Correlations between the separate peg boards and M.A. are reported by Goodenough as follows:

Board	Age 2	Age 3	Age 4
A	.46	.29	.37
B	.47	.39	.28
C		.42	.40
D		.42	.48

The average correlation of the four peg boards with Kuhlmann-Binet Mental Age was .51 for single-year age groups. These tests involve general intelligence as defined by the Binet M.A., but it is clear that they also draw upon other functions to a greater degree. Stutsman states that the peg-board performance gives valuable clues to the personality of the child (48), while Goodenough is of the opinion that the tests are related to mechanical aptitude. The tests are easy to give and to score and there is apparently no difference in the peg-board ability of children from different social and economic levels. Age norms for boys and girls are given for the Peg Board Tests by Goodenough (22) and by Hallowell (25).

### 5. Hand Test

This test was devised by Thurstone (45) and has been used with some modification by Squires (47). It is a sub-test in the Princeton Universal Scale of Performance Tests. The material consists of a series of drawings made of the right and left hand in various positions. The subject is required to indicate whether the right or left hand is represented by the drawing. Thurstone's original test contained forty-nine drawings, while Squires' series had thirty-seven, seven of which were for practice. Scoring is in terms of the number of hands correctly indicated; and the time taken to complete the test is also recorded. Thurstone has given the test to 238 men and 114 women college students, and percentile norms for this group are given by Bronner, Healy, *et al.* (4).

Squires has tested fifty seniors at Princeton University with the Hand Test and has correlated these scores with general intelligence tests. The correlation between number right on the Hand Test and intelligence test scores was .10; and when time was used as the score the correlation was —.24. This test would probably be useful if included in a battery of tests designed to measure spatial perception.

## 6. Minnesota Spatial Relations Test

The Minnesota Spatial Relations Test (36) has been designed especially to measure space perception. This test is a modification and extension of a test devised by Link (31). Link's test consisted of two boards in which depressions of various sizes and shapes were cut. Into these depressions blocks were to be fitted by the subject. Some pieces were very different in shape, while others were almost alike, differing slightly in size only. Link's test was used in the preliminary work of the Minnesota study of mechanical ability. Three trials were given for each board and the performance scored in terms of time and errors. An error was defined as an attempt to fit a block into the wrong depression. In a group of 217 junior-high-school boys the reliability of this test (first trial omitted) was found to be .60. Its correlation with the combined shop criterion score (actual performance in shop work) was .36 for the two boards together and .44 for the longer of the two boards. In order to secure higher reliability the Minnesota workers increased the number of boards from two, each with nineteen items, to four, each with fifty-seven items. This lengthening of the test raised the reliability to .84 and its correlation with quality of work in shop courses to .53.

The intercorrelations of this test with some of the other tests used in the Minnesota study are given in the following table (36).

N = 100 SEVENTH- AND EIGHTH-GRADE BOYS	
Correlation of Spatial Relations Test with	r
Card Sorting . . . . .	.27
Packing Blocks . . . . .	.42
Paper Form Board . . . . .	.72
Steadiness . . . . .	— .01
Stenquist Picture Test I . . . . .	.48
Minnesota Assembly Test . . . . .	.63

The highest correlation is that of the Minnesota Spatial Relations Test and the Paper Form Board Test, which is clearly a measure of complex spatial relationships. The positive correlation of the Spatial Relations Test with Card Sorting and Block Packing probably arises from a common manipulative element in these tests. The correlation of the Spatial Relations Test with Steadiness is zero, as might be expected from the diverse character of the two measures.

The relationship of the Minnesota Spatial Relations Test to measures of general intelligence is low, the correlation with Otis I.Q. being .18. A somewhat surprising result obtained by the Minnesota investigators was the superiority of seventh-grade girls over seventh-

grade boys in the Spatial Relations Test, since boys are usually superior in manipulative motor tests. For university students the superiority was in favor of the men. Engineering students were slightly inferior in this test to students in the college of liberal arts. This rather surprising result indicates the probability that engineering students do not possess greater native ability to visualize spatial relations than do academic students.

### 7. Wiggly Block Test

This test is a combined form-board and construction test. It was devised by O'Connor (34), who called it a "work sample" rather than a test. The test as shown in Figure 16 consists of a block of wood sawed into nine wavy pieces. When assembled the block contains three layers of pieces, three in each layer. In giving the test the block is taken apart and the pieces mixed up in certain prescribed ways. The subjects are then told to reconstruct it. The reliability of the test as reported by O'Connor is .36 (first trial against second trial). This is, of course, extremely low, but O'Connor seems to regard the test as fairly reliable since its correlation with another form board, devised by Kent (34), was .76. It should be pointed out that such a correlation is not a measure of reliability, but indicates simply that the Wiggly Block Test measures much the same function as other form boards. O'Connor assigns a weighted score to each trial with the Wiggly Block Test, in order to make succeeding trials comparable to the first. Three trials are given and scoring is in terms of time.

It is difficult to say just what the Wiggly Block Test measures. O'Connor considers it to be a measure of the ability "to visualize three-dimensional structure," and has used the test as a measure of mechanical aptitude. Its correlation with other form boards, however, makes it seem probable that the test is measuring discrimination and manipulation, primarily, and for this reason it was included in this chapter rather than in Chapter II.<sup>1</sup> O'Connor has given the Wiggly Block Test to about 4,000 men, including mechanics, engineers, draftsmen, machine tenders, *etc.* It was found that 82 per cent. of the engineers scored above the median for the whole group; that 75 per cent. of the draftsmen, and 71 per cent. of the mechanics also made scores above the median. On the other hand, the majority of the machine tenders were below the median and none of them fell

<sup>1</sup>Part One. Chapters in Part Two are indicated by italics.

in the upper 25 per cent. of the distribution. O'Connor concludes that the ability measured by the Wiggly Block Test is not needed in

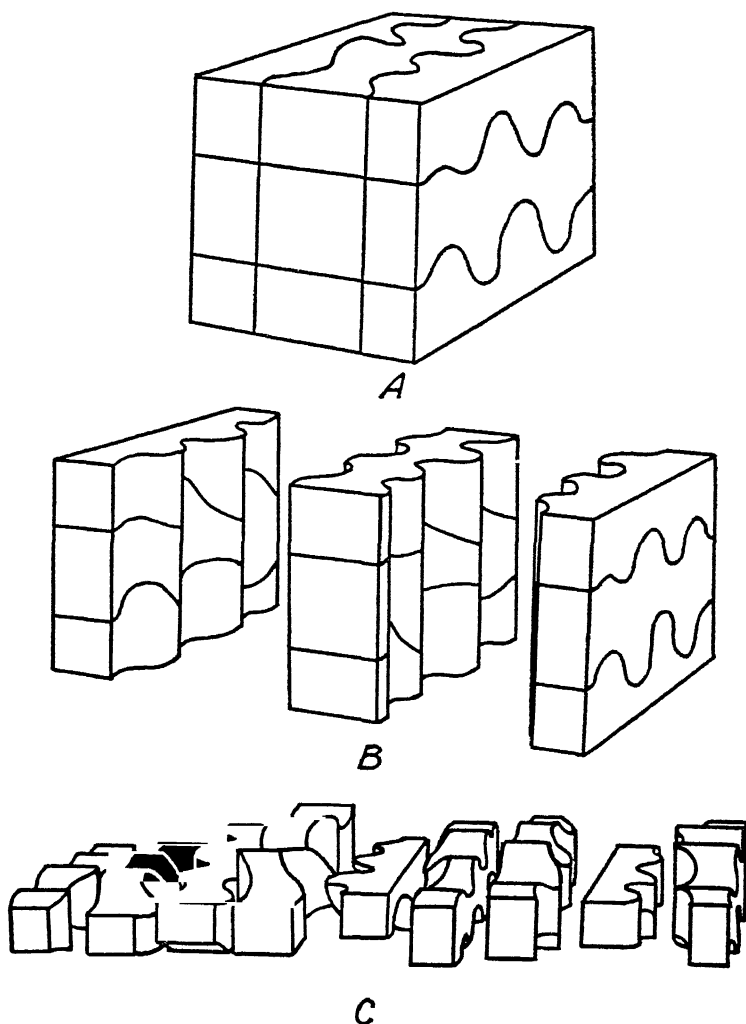


Figure 16.—O'CONNOR WIGGLY BLOCK TEST

machine tending. All machine tenders who graded A on the test were later advanced to the position of mechanic.

O'Connor lists a number of occupations requiring mechanical aptitude which he believes may be successfully undertaken by those who grade A or B in the Wiggly Block Test. The assumption made

is that these types of work depend upon the same abilities which are measured by the test. Evidence for the validity of the test as a measure of mechanical ability is not sufficiently complete to justify the author's conclusion; but practical results do indicate that there is enough real value in the test to justify its further study. Detailed instructions for administration of the test are given by O'Connor (34).

### 8. Minnesota Paper Form Board

The Paper Form Board is one of the Minnesota Mechanical Ability Tests. It is included in this chapter because it deals directly with form and with two-dimensional space perception (36). Originally the test consisted of thirty problems resembling those in Test 7 of the Army Beta examination (59). Each problem consists of a large geometrical figure into which a number of smaller figures are to be fitted. This is accomplished by drawing lines in the large figure so as to reproduce the smaller ones, and fill up completely the space in the large figure. (See Figure 17.)

The test consists of two forms of fifty-six items each, the reliability of a single form being .90. Scoring is in terms of the number of correct solutions. The correlation of the Paper Form Board with the quality criterion of mechanical ability employed in the Minnesota experiments was .52; with a manual-training information test .57; and with a combination of quality of work and information .65. These correlations show that the Paper Form Board Test should be useful as a member of a battery of tests designed to measure mechanical ability.

The correlations of the Paper Form Board with other tests used in the Minnesota experiments throw some light upon the functions probably measured by this test. The correlations of the Paper Form Board with Card Sorting (.16) and with Packing Blocks (.17) are to be contrasted with its correlations with Spatial Relations (.72) and with the Minnesota Assembling Test (.53). Card Sorting and Packing Blocks have correlations with the Spatial Relations Test (*viz.*, .27 and .42), which are higher than the correlations of these tests with the Paper Form Board. This is probably due to the fact that the Paper Form Board does not require actual manipulation and motor coördination as does the Spatial Relations Test. At the same time, however, there is a substantial community of function

## Minnesota Paper Form Board Test Series A.

Name . . . . . Month and Year of Birth . . . . .

School . . . . . Class . . . . .

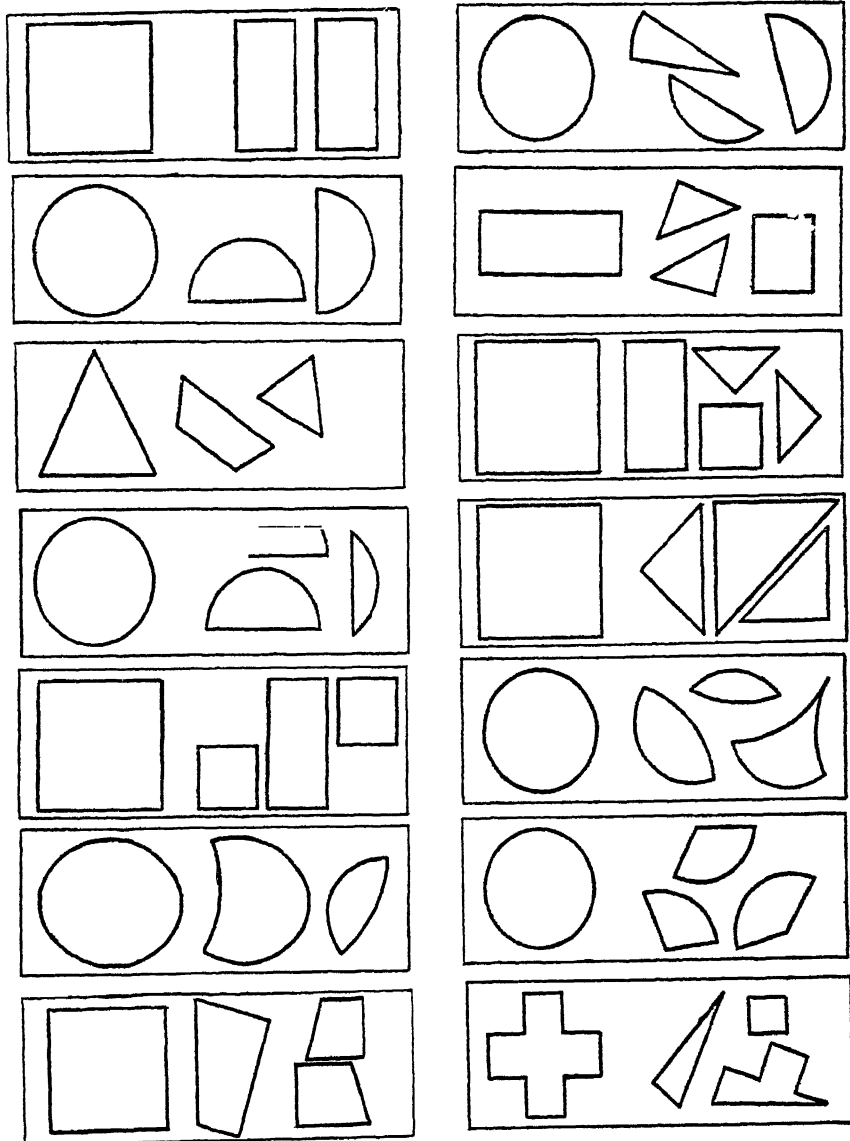
MECHANICAL ABILITIES PROJECT  
DEPARTMENT OF PSYCHOLOGY  
UNIVERSITY OF MINNESOTA

Figure 17.—MINNESOTA PAPER FORM BOARD TEST

(Specimen page)

between the Paper Form Board and the Spatial Relations Test, although the one deals with abstract and the other with concrete relations.

The correlation of the Paper Form Board Test with I.Q. as determined from the Otis S.A. Test was .53 in a group of 100 boys. This is to be contrasted with the correlation of .18 between the more manipulative Spatial Relations Test and Otis I.Q. in the same group. The substantial relationship between Otis I.Q. and the Paper Form Board is owing probably to the fact that the Otis Test contains many elements which involve numerical and spatial relations. With a more verbal test of general intelligence the correlation of the Paper Form Board would be extremely low. Anastasi (1), for example, obtained a correlation of .07 between the Paper Form Board and a difficult vocabulary test in a group of 225 college men. In the same group the correlation of the Paper Form Board with arithmetic reasoning was .30.

There is a slight superiority of men and boys over women and girls in the Paper Form Board Test. This difference is found in the seventh grade as well as in the university, although it is not large enough to be statistically significant.

### 9. Witmer Cylinder Test

This is a performance test which calls for form and space perception of a high order. The test (Figure 18) consists of a circular block of wood in the outer edge of which are cut eighteen depressions of different depths and diameters. The subject is asked to fit cylinders into these depressions. Since these cylinders differ in height and in diameter, the test is more difficult than the ordinary form board. Witmer is of the opinion that the test measures planfulness, analytic ability, and sustained attention.

The Cylinder Test was standardized by Paschal (35) upon a group of 1,221 men and boys and 1,009 women and girls. Paschal has stressed the usefulness of the test in the study of qualitative differences in performance. He writes, "qualitatively the first trial tests ability to make an adaptation to a new problem, and is sufficiently complicated as to permit the differentiation of subjects especially in the years of childhood." The test has found wide use in psychological clinics.

Paschal found a steady decrease in time required to complete the

test with increase in age up to adulthood. According to this author the shortest of three trials is the best index of "psychomotor capacity." In support of this conclusion, he points to the fact that such scores give higher correlations with proficiency in shop trades than do other methods of scoring. Some of the correlations reported with the Cylinder Test are decidedly high, as, for example, the correlation of .69 between the Cylinder Test and tailoring, and .72 between the Cylinder Test and shoe-making. Males are consistently superior to females on this test.

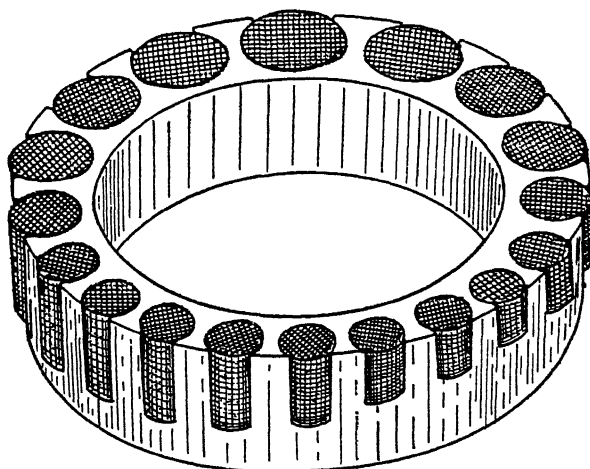


Figure 18.—WITMER CYLINDER TEST

While the Cylinder Test was not intended to be primarily a measure of abstract intelligence, its correlation with intelligence tests is fairly high. Johnson and Schriefer (28), in a group of eighty-six children three to ten years old, report a correlation of the Cylinder Test with Stanford-Binet M.A. of .58; and with Pintner-Paterson Median Mental Age of .67. These correlations are probably somewhat "inflated" because of the wide age range, but they serve nevertheless to demonstrate considerable overlap in the abilities measured by the Cylinder Test and the general intelligence batteries.

There are many tests of perception and attention which have not been included in the present chapter. We have tried to present, however, tests which are useful in the comparison of individuals, and in the study of the relation of perception and attention to other types of performance. Precise measurements of the fineness of discrimination (using weights, brightnesses, areas, tones, and so forth) by



means of the psycho-physical methods have not been included. Such measurements clearly involve perception and attention, but they are adapted to laboratory use rather than to the broad study of individual differences. Also, performance tests, such as form boards, and picture-completion tests, have been omitted, although such tests clearly measure perception and attention. These tests, however, have usually been included in batteries designed to measure general ability as shown in concrete performance. Hence they will be found in Chapter II, which deals with non-language and performance tests of general intelligence. Many individual tests of perception have been described by Schieffelin and Schwesinger (45), by Bronner, Healy, *et al.* (4), by Whipple (55), by Stutsman (48), by Pintner-Paterson (38) and by Whitley (56).

The tests described in this chapter may be purchased from the C. H. Stoelting Co., Chicago, Illinois, or the Marietta Apparatus Co., Marietta, Ohio.

#### BIBLIOGRAPHY

1. ANASTASI, A., "A Group Factor in Immediate Memory," *Archives Psychology*, No. 120, 1930.
2. ATKINSON, W. R., "The Relation of Intelligence and of Mechanical Speeds to the Various Stages of Learning," *Journal Experimental Psychology*, 12:89-112, 1929.
3. BALDWIN, B. T., AND STECHER, L. I., *The Psychology of the Pre-school Child*, D. Appleton & Co., New York, 1924.
4. BRONNER, A., HEALY, W., LOWE, G., AND SHIMBERG, M., *A Manual of Individual Mental Tests and Testing*, Little, Brown & Co., Boston, 1927.
5. BROWN, W., *The Essentials of Mental Measurement*, Cambridge, 1911.
6. BUSWELL, G. T., *Fundamental Reading Habits. A Study of Their Development*, University of Chicago Press, Chicago, 1922.
7. CALFEE, M., "College Freshmen and Four General Intelligence Tests," *Journal Educational Psychology*, 4:223-231, 1913.
8. CAROTHERS, F. E., "Psychological Examinations of College Students," *Archives Psychology*, No. 46, 1921.
9. CATTELL, J. McKEEN, "The Inertia of the Eye and Brain," *Brain*, 8:295-312, 1885-1886.
10. CORNELL, C. B., "A Graduated Scale for Determining Mental Age," *Journal Educational Psychology*, 8:539-550, 1917.
11. CRAFTS, L. W., "Routine of Varying Practice as Preparation for Adjustment to a New Situation," *Archives Psychology*, No. 91, 1927.
12. DEWEERDT, E. H., "A Study of the Improvability of Fifth-grade School Children in Certain Mental Functions," *Journal Educational Psychology*, 18:547-557, 1927.

13. DALLENBACH, K. M., "The Effect of Practice upon Visual Apprehension in School Children," *Journal Educational Psychology*, 5:321-334. 387-404, 1914.
14. DALLENBACH, K. M., "The Effect of Practice upon Visual Apprehension in the Feeble-minded," *Journal Educational Psychology*, 10:61-82. 1919.
15. DEARBORN, W. F., "Psychology of Reading," *Archives Philosophy, Psychology, and Scientific Method*, No. 1, 1906.
16. DEWEY, E., CHILD, E., AND RUMMLER, B., *Methods and Results of Testing School Children*, E. P. Dutton & Company. New York, 1920.
17. FOSTER, W. S., "The Effect of Practice upon Visualizing and upon the Reproduction of Visual Impressions," *Journal Educational Psychology*, 2:11-22. 1911.
18. GARRETT, H. E., AND LEMMON, V. W., "Analysis of Several Well-known Tests," *Journal Applied Psychology*, 8:424-438, 1924.
19. GARRISON, K. C., "An Investigation of Some Simple Speed Activities," *Journal Applied Psychology*, 13:167-172, 1929.
20. GATES, A. I., "A Critique of Methods of Estimating and Measuring the Transfer of Training," *Journal Educational Psychology*, 15:545-558. 1924.
21. GATES, A. I., "Variations in Efficiency During the Day, Together with Practice Effects, Sex Differences, and Correlations," *University of California Publications in Psychology*, No. 1, 2, 1916.
22. GOODENOUGH, F. L., "The Reliability and Validity of the Wallin Peg Boards," *Psychological Clinic*, 16:199-215, 1927.
23. GRAY, C. T., *Deficiencies in Reading Ability*, D. C. Heath and Company, Boston, 1922.
24. GRAY, C. T., *Types of Reading Ability as Exhibited Through Tests and Laboratory Experiments*, Supplementary Education Monographs, University of Chicago Press, Chicago, No. 5, 1, 1917.
25. HALLOWELL, D. K., "Mental Tests for Pre-school Children," *Psychological Clinic*, 16:235-276, 1928.
26. HERTZBERG, O. E., "Relationship of Motor Ability to the Intelligence of Kindergarten Children," *Journal Educational Psychology*, 20:507-519, 1929.
27. HOFFMAN, J., "Experimentelle-psychologische Untersuchungen über Leseleistungen von Schulkindern," *Archiv für die gesamte Psychologie*, 58:325-388, 1927.
28. JOHNSON, B., AND SCHRIEFER, LOUISE, "A Comparison of Mental Age Scores Obtained by Performance Tests and the Stanford Revision of the Binet-Simon Scale," *Journal Educational Psychology*, 13:408-417, 1922.
29. JONES, H. E., "Dextrality as a Function of Age," *Journal Experimental Psychology*, 14:125-143, 1931.
30. KINCAID, M., "A Study of Individual Differences in Learning," *Psychological Review*, 32:34-54, 1925.

31. LINK, H. C., *Employment Psychology*, The Macmillan Company, New York, 1920.
32. MCCALL, W. A., *Correlation of Some Psychological and Educational Measurements*, Teachers College, Columbia University, Contributions to Education, No. 79, 1916.
33. O'BRIEN, J. A., *Reading—Its Psychology and Pedagogy*, The Century Company, New York, 1926.
34. O'CONNOR, J., *Born That Way*, The Williams and Wilkins Company, Baltimore, 1928.
35. PASCHAL, F. C., *The Witmer Cylinder Test*, The Hershey Press, Hershey, Pa., 1918.
36. PATERSON, D. G., ELLIOTT, R. M., *et al.*, *Minnesota Mechanical Ability Tests*, University of Minnesota Press, Minneapolis, 1930.
37. PETERSON, J., AND BARLOW, M. C., "The Effects of Practice on Individual Differences," *27th Yearbook, National Society Study Education*, Part II, 211–230, 1928.
38. PINTNER, R., AND PATERSON, D. G., *A Scale of Performance Tests*, D. Appleton & Company, New York, 1925.
39. PYLE, W. H., *The Examination of School Children*, The Macmillan Company, New York, 1913.
40. PYLE, W. H., *Laboratory Manual in the Psychology of Learning*, Warwick and York, Inc., Baltimore, 1923.
41. PYLE, W. H., AND SNADDEN, G. H., "An Experimental Study of Bright and Dull High School Pupils," *Journal Educational Psychology*, 20:262–269, 1929.
42. RANSCHBURG, P., "Die Leseschwache und Rechenschwache der Schulkinder im Lichte des Experiments," *Zwanglose Abh. aus den Grenzgebieten der Pädagogik und Medizin*, Heft 7, Berlin, 1916.
43. RUCH, G. M., "Correlations of the Initial and Final Capacities in Learning," *Journal Experimental Psychology*, 6:344–356, 1923.
44. RUCH, G. M., "The Influence of the Factor of Intelligence on the Form of the Learning Curve," *Psychological Monographs*, No. 7, 34, 1925.
45. SCHIEFFELIN, B., AND SCHWESINGER, G. C., *Mental Tests and Heredity*, The Galton Publishing Company, New York, 1930.
46. SOMMERVILLE, R., "Physical, Motor, and Sensory Traits," *Archives Psychology*, No. 75, 1924.
47. SQUIRES, P. C., *A Universal Scale of Individual Performance Tests*, Princeton University Press, Princeton, 1926.
48. STUTSMAN, R., *Mental Measurement of Pre-school Children*, World Book Company, Yonkers, 1931.
49. TINKER, M. A., "Visual Apprehension and Perception in Reading," *Psychological Bulletin*, 26:223–240, 1929.
50. TINKER, M. A., IMM, A. J., AND SWANSON, C. A., "Card-sorting as a Measure of Learning and Serial Action," *Journal Experimental Psychology*, 15:206–211, 1932.
51. VICKERY, K., "The Effects of Change of Work on the Work Decrement," *Journal Experimental Psychology*, 14:218–241, 1931.

52. WALLIN, J. E. W., "The Peg Form Boards," *Psychological Clinic*, 12: 40-53, 1918.
53. WELLMAN, B., "The Development of Motor Coördination in Young Children," *University of Iowa Studies in Child Welfare*, No. 4, 3, 1926.
54. WHIPPLE, G. M., "The Effect of Practise upon the Range of Visual Attention and of Visual Apprehension," *Journal Educational Psychology*, 1:249-262, 1910.
55. WHIPPLE, G. M., *Manual of Mental and Physical Tests, Simpler Processes*, Warwick and York, Baltimore, 1914.
56. WHITLEY, M. T., "An Empirical Study of Certain Tests for Individual Differences," *Archives Psychology*, No. 19, 1911.
57. WOODWORTH, R. S., AND WELLS, F. L., "Association Tests," *Psychological Monographs*, No. 57, 13, 1911.
58. WOOLLEY, H. T., "A New Scale of Mental and Physical Measurements for Adolescents, and Some of Its Uses," *Journal Educational Psychology*, 6:521-550, 1915.
59. YOAKUM, C. S., AND YERKES, R. M., *Army Mental Tests*, Henry Holt & Company, Inc., New York, 1920.

## CHAPTER IV

### TESTS OF LEARNING, ASSOCIATION AND MEMORY

THE terms learning, association and memory describe interrelated activities which overlap so frequently that their separation for purposes of testing is often more a matter of convenience than of reality. Differences in the object or purpose of tests designed to measure these functions, however, can usually be drawn fairly definitely. Tests of learning attempt to measure S's ability to form new associations and to acquire new facts through active effort. Tests of association are concerned with the efficiency, *i.e.*, the speed and accuracy, with which associations already formed can be reproduced or made to function. Such tests, therefore, deal with facts which have already been learned. Memory tests, in contradistinction to tests of learning and association, are measures of how accurately material which has once been presented can be recalled or recognized either immediately or at some future time.

Certain representative tests which have been designed primarily to measure learning or association or memory will be presented in this chapter. In most instances the distinctions among such measures are fairly closely adhered to, and the primary function of the test can be rather readily recognized. Oftentimes, however, it is almost impossible to label a test definitely as a measure of memory or association or learning. In such cases, the test has been classified somewhat arbitrarily, perhaps, where it seemed to fit best.

#### TESTS OF LEARNING

In the following sections, tests of learning have been arranged under two headings. In the first category are those tests which require the formation of designated associations or connections among words, numbers or other symbols (*i.e.*, "ideas"). These tests are sometimes described as tests of *ideational learning*. In the second category are tests in which associations between movements, as well as between movements and ideas, are required. These are often

called tests of *sensory-motor learning*, since in their performance sense perception as well as speed and precision of movement play important rôles.

The tasks set by learning tests may seem at first glance to be so different from learning mathematics or learning how to play chess or how to run an automobile that the student may well wonder why psychologists have ever devised such measures. The reason is that the processes involved in the performance of these tests are basically the same as those operative in many everyday learning situations. Learning involves essentially the orderly linking together of ideas or movements, or both. By simplifying this process and by cutting away to a large degree the factor of experience, the test of learning enables one the more easily to make comparisons between individuals or between groups of widely different attainments, or of widely different backgrounds.

## I. IDEATIONAL TESTS

### 1. Substitution Tests

In these tests S is required to substitute one set of characters (numbers, letters, geometrical forms, *etc.*) for another set in accordance with the instructions given in a "key." The associations are not learned all at once, but are gradually acquired as a result of active repetition and effort. A familiar form of the substitution test (82), shown in Figure 19, requires the placing of digits in geometrical forms. Each of the five geometrical forms printed at the top of the sheet contains a single digit. Below this row of numbered forms (called the "key") are the same five geometrical forms repeated twenty times each (making 100 in all) and arranged in chance order. These 100 forms are unnumbered. S's task is to write within each of the unnumbered forms the digit found in the corresponding geometrical form in the key. S learns as he goes along, *i.e.*, he gradually comes to link each form with its appropriate number until finally he no longer has to refer to the key. In other forms of substitution learning, S is required to replace digits by letters, letters by symbols, or to write a paragraph in code. A code test has been placed in the Stanford-Binet scale at the adult level, and a modified code test was employed by Ruch (62) and by Dearborn and Lincoln (16). Foster and Tinker (18) give a list of thirty Turkish-English equivalents which

may be used as a substitution learning test. Several ingenious learning tests which involve the association of symbols of various sorts will be found in Bronner, Healy, *et al.* (7).

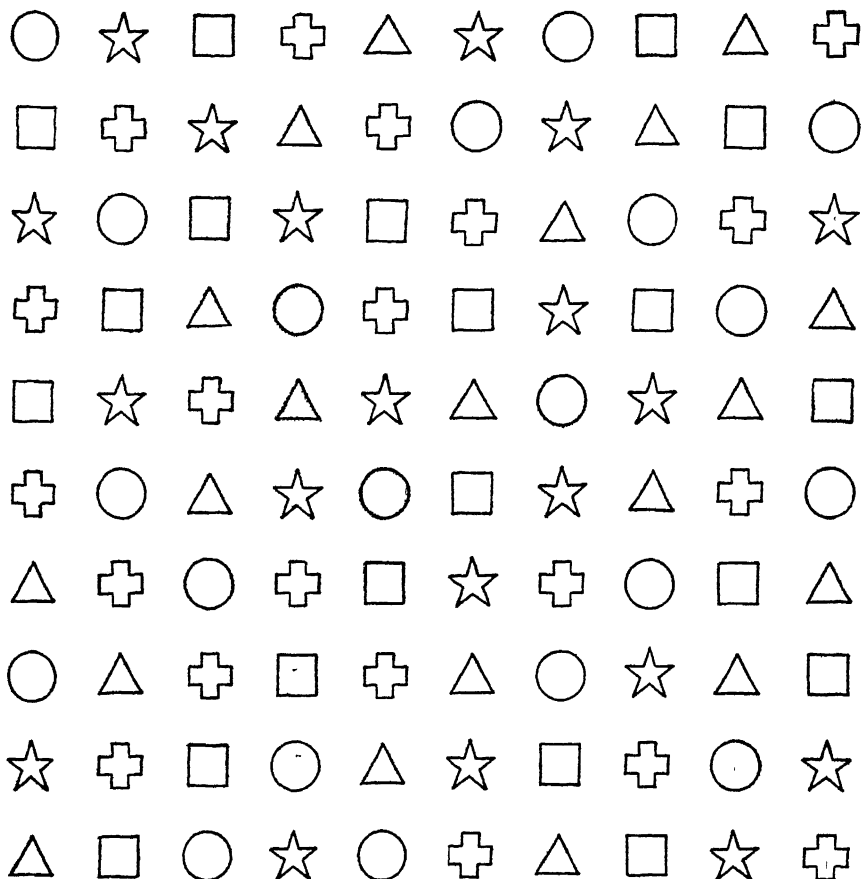


Figure 19.—WOODWORTH-WELLS DIGIT-SYMBOL SUBSTITUTION TEST

In a somewhat different kind of substitution test devised by Thurstone (7), S is given a list of 600 letters, arranged in thirty columns, each letter followed by a dash. These letters are the initials of twenty words printed at the top of the sheet. S's task is to write the last letter

of the correct word next to each initial letter. Learning is measured by the quickness with which the subject is able to make the proper combinations as he goes through the sheet.

## 2. Rational Learning Test

The rational learning test was devised by Peterson (52), who considers it to be a test of intelligence as well as a test of learning. The student is required to learn associations between certain designated numbers and letters as quickly as possible. The task is as follows. The letters A to J, inclusive, are numbered in random order from 1 to 10, the number of each letter being unknown to S. The experimenter calls out the first letter A, and instructs S to guess numbers until he guesses the right one for A. The experimenter then says "right," and calls out the next letter, B. S again guesses numbers until he hits upon the "right" one for B. The process continues until S gets each number right twice in succession through the series from A to J. S's performance is judged by (a) the total time taken, (b) the number of errors he makes, and (c) the number of repetitions from A to J required. A sample of the experimental set-up is shown below.

A	B	C	D	E	F	G	H	I	J
6	2	8	5	9	10	4	7	3	1

S is left free to adopt whatever method of learning he finds best. The nature of the learning task is indicated to some degree by the types of errors made by the subjects. Peterson distinguishes three classes of errors: (a) Logical errors, or the repetition of numbers which have already been used for earlier letters of the series and which could not, according to the conditions of the experiment, be correct; (b) perseverative errors, or repetition of numbers already guessed incorrectly for the letter in question, before the correct number is found; (c) unclassified errors, which include all other types. An incorrect response may be given three error counts, since it may fall into all three classes of errors. This test may be made more difficult by increasing the length of the letter series.

## 3. Modified Form of the Rational Learning Test

An apparatus for administering the rational learning test has been designed by Haught (29). This apparatus consists of a board twenty inches square, through which are put 100 bolts arranged in ten rows, ten bolts in each row. The rows are lettered from A to J, and the bolts in each row are numbered from 1 to 10. One bolt in each



row is connected electrically so that a bell rings when the bolt is touched by a metal stylus. S begins with row A and finds by successive trials the bolt in that row that will ring the bell. He then goes on to the next row, and so on through to row J. S repeats the process until he is able to go from row A to row J twice in succession without error, *i.e.*, without touching any bolts except those which ring the bell.

#### 4. The Mental Maze

The mental maze, also designed by Peterson (50), is in principle closely allied to the rational learning test. S is shown a drawing like that in Figure 20 and is given the following instructions. "This draw-

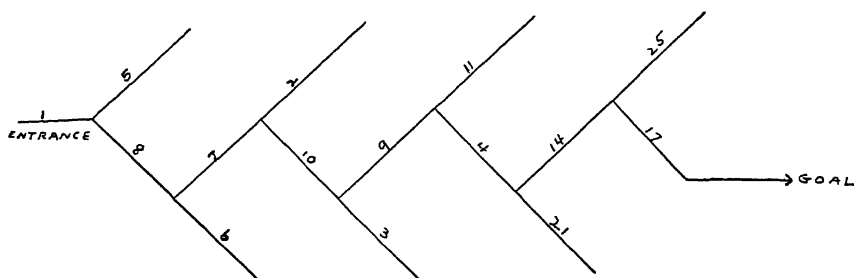


Figure 20—PETERSON MENTAL MAZE

ing is a drawing of a maze. Let us suppose that you are at the beginning, 1, and want to get through to the goal. Where would you go if the lines are paths? We shall see how this maze works out and then you will go through another just like it. Let us begin at 1. I shall call out two numbers for the first two paths and you are to choose one of them. Then I shall call out two more numbers and you must choose one of them, and so on until you reach the goal." The subject learns that choosing the "right" number gets him farther along in the maze, while choosing the "wrong" number is counted an error. The procedure may be varied by using words instead of numbers, S being required to guess the "right" word of the pair. When S is able to reach the goal twice in succession without error, the task is finished. The test is scored in terms of minutes taken in learning, number of errors made, and number of repetitions required. A maximum time limit of forty minutes was set by Peterson.

#### 5. Results Obtained with Ideational Learning Tests

(a) *The Relationship of Ideational Learning to General Intelli-*

*gence and to Other Capacities.*—Ideational learning tests have been studied in connection with many other functions, but perhaps the relationship of such tests to measures of general intelligence has been of greatest interest to investigators (14). In the case of children, the relation of learning tests to measures of general intelligence is fairly high, since to children such tests offer a real task. Ideational learning increases directly with age. The improvement in substitution learning with age, for example, is clearly shown in Table XV (p. 100); and Willoughby (79), who administered a digit-symbol test to 300 subjects ranging in age from six to sixty-eight years, found a sharp increase in score up to age twenty, followed by a gradual decline thereafter.

Willoughby (79) has reported a correlation of .66 between I.Q. and substitution learning in a group of sixty-four children, ages  $12\frac{1}{2}$  to  $13\frac{1}{2}$ . The learning test was the digit-symbol test in Army Beta, while the I.Q.'s were computed from a battery of nine tests which included opposites, arithmetic reasoning, vocabulary, analogies, *etc.* Willoughby's correlation of .66 is somewhat too high, since the battery of tests from which he obtained the I.Q.'s included the substitution test; but even allowing for this spurious factor, the relationship would still be substantial.

Ruch (62) administered a code substitution test to sixty-six children in grades 7, 8 and 9, the average chronological age of the group being fourteen years. The test consisted of translating a chapter of *Oliver Twist* from code into English; and another chapter from English into code. Daily practice for ten minutes on each kind of translation was given over a period of ten days. These daily scores were then correlated with the children's M.A.'s, obtained from the Stanford-Binet. The results were as follows: for translation into English the correlations with I.Q. ranged from .61 to .70; for translation into code from .51 to .67. The size of these coefficients was not changed appreciably when age variability was allowed for by partial correlation. Peterson and Lanier (55) have reported a correlation of .37 between a group form of the Stanford-Binet and the rational learning test in a group of 107 twelve-year-old white children. The scores on the rational learning test were based upon time, repetitions and number of errors. In the same group the correlation of the rational learning test and scores in the Myers Mental Measure was .53.

The correlations between ideational learning tests and general in-

telligence scores are considerably lower for adults than for children. A correlation of .01, for instance, has been reported by Garrison (22) between the Otis Self-Administering Test and a number-letter substitution test, in a group of sixty college students. Peterson and Lanier (55), who gave a digit-symbol test to 130 white and seventy-seven Negro college students, obtained correlations of .22 for the whites and .19 for the Negroes between this test and the Otis Self-Administering Test. Garrett (19) obtained a correlation of .09 between the Thorndike Intelligence Examination and the digit-symbol test in a group of 158 college men. In the same group, the correlation of the Thorndike Intelligence Test with a Turkish-English vocabulary substitution test was .37, and with a code-learning test .31. Fisher (17) administered the Army Alpha along with a test in which digits were to be substituted for five geometrical forms, to eighty-three men and boys and seventy-four women and girls, selected from a larger group. The correlations of substitution-learning and Army Alpha were .34 for the males and .20 for the females. These  $r$ 's are probably somewhat high because of the wide age range.

The rational learning test (53) is more closely related to general intelligence test scores in the case of adults than is the simple substitution test. This is probably owing to the fact that rational learning requires more concentrated effort and is a more difficult task. Garrison (21) has reported a correlation of .51 between Army Alpha and the rational learning test in a group of forty college men. Haught (29), who gave the modified rational learning test to seventy-four college freshmen and sophomores, obtained a correlation of .47 between error scores in rational learning and Stanford-Binet M.A.

In a group of 119 white children, Peterson and Lanier (55) obtained a correlation of .27 between rational learning *time* and the mental maze *time*; while in a group of eighty-six Negroes the correlation of the same two tests was .33. Though similar in form, these tests seem to be fairly specific in function. These authors made comparisons of white and Negro children upon the rational learning test in Nashville, Chicago and New York. An interesting finding was the fact that the white children, although superior to the Negroes in the first two cities, were inferior to them in the third, thus suggesting strongly the operation of selective factors. The correlations of the Stanford-Binet with the rational learning test ranged from .33 to .46

in the white and Negro groups, the differences between the correlations found in the two groups being slight.

(b) *Effects of Practice*.—An interesting study of the change in the character of the substitution learning task with practice has been reported by Atkinson (4). Over a period of twenty days, thirty college students were required to practice on the following tests: a number-code substitution test; a speed of writing test, in which the subject wrote all the numbers he could in one minute; silent counting from 1 on, until time was called. The correlations between substitution and speed of writing rose from .30 in the first trial to .83 in the twentieth trial. This indicates that, with continued practice, substitution learning (for adults at least) becomes largely a matter of speed of writing. Atkinson also correlated the daily record of his subjects in the number-code test with Army Alpha scores. The correlation between Alpha and the first day's practice in the code was .70; while the correlation between Alpha and the twentieth practice period in substitution was .49. This drop in correlation with practice in substitution probably resulted from the fact that with constant repetition the code test becomes more and more a matter of mechanical writing speed.

(c) *Reliability*.—The reliability of ideational learning tests is, in general, quite high. In the Minnesota Study (49) the self-correlation of the digit-symbol test was .78 in a group of 217 junior-high-school boys. Carothers (11) has reported a reliability of .70 for the Woodworth-Wells number-form substitution test in a group of forty-five freshmen women. A higher reliability for the Whipple digit-symbol test was obtained by Lemmon (39), who found a retest reliability of .95 in a group of ninety-two college men. Garrett (19) obtained reliabilities of .95, .91 and .85 for the digit-symbol, Turkish-English vocabulary and code-learning tests, respectively, in a group of 158 college men.

The reliability of the rational learning test was investigated by Peterson and Lanier (55). In a group of forty-nine college students who were given two forms of an eight-letter test, the reliability of the time scores was .75; of the repetition scores .60, and of the error scores .70. The reliability of the whole test when time, repetition and errors were combined with equal weight was .67. In another study, Peterson (51) has reported the reliability of the error scores in rational learning to be .92 in a group of 100 college students.

## Test 1. Digit-Symbol Substitution Test

*Material:* Whipple's digit-symbol<sup>1</sup> test blanks; stop-watch. Test blanks may be procured from the C. H. Stoelting Company, Chicago, Illinois.

*Method.* This test may be used either for individual or for group testing. Place a blank face down before S and instruct him as follows: "At the top of the sheet are nine circles, each circle containing a number and a figure. Below this key you will find rows of numbers. The first row of numbers, for instance, is 8 4 9 7 6. The first number is 8, and in the circle above which contains an 8 you will find an X sign. Write an X sign in the first square. (If the test is being given to children E should demonstrate on the blackboard.) The next number is 4. In the circle containing a 4 you will find an O. Put this O in the next square and fill in the other three squares in the same way by referring back to the key. Take the rows in order and work down the page. When you finish the first half go on to the second half. Work as rapidly as you can without making mistakes and do not skip any. When I say *go* begin at once, and stop when I say *stop*." Five minutes should be allowed between the *go* and *stop* signals.

*Record:* The score is the number of correct substitutions made in five minutes. Errors may be taken into account by subtracting 1 to 2 points from the score for each error. If the test is given by the work-limit method, S is allowed to finish and his score is the time taken to complete.

*Norms:* Table XV gives the average (smoothed) scores made by children in the digit-symbol test. The time allowance was five minutes. These data, taken from Pyle, are based upon nearly 8,000 cases.

TABLE XV  
NUMBER OF SUBSTITUTIONS MADE IN FIVE MINUTES (DIGIT-SYMBOL  
TEST)  
(from Pyle [57])

Age	Boys		Girls	
	N	Score	N	Score
8 .	260	37 5	265	42 5
9 .	355	47 5	416	56 0
10 . .	488	57 5	465	68 2
11 . .	519	67 5	461	80 2
12 . .	473	77 5	524	91 6
13 . .	476	87 5	504	102 3
14 . .	398	97 5	440	112 3
15 . .	332	107 5	374	122 3
16 . .	245	117 5	285	133 3
17 . .	131	127 5	213	138 7
18 . .	52	137 0	86	142 0

In the Columbia Laboratory the following results have been obtained with the digit-symbol test. S's were men and women; and the time limit was four minutes N = 52, Average = 146.93 and S.D. = 30.09.

<sup>1</sup> The symbol-digit blank may also be used.

## II. SENSORY-MOTOR LEARNING

## 1. Maze Tracing

Paper and pencil mazes have been widely employed as measures of general intelligence and of learning capacity (56) and for studying hand-eye coördination and voluntary control (77). In the experimental study of human learning the mazes employed have been usually stylus or finger mazes, in which the subject (usually blindfolded) is required to trace through the grooves of the maze with his finger or with a stylus. (See Figure 21.) The subject is given trial after trial

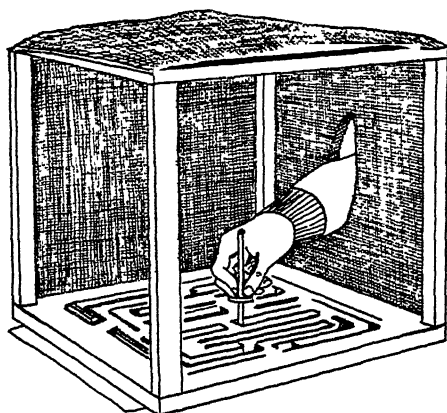


Figure 21.—STYLUS MAZE (KLINE)  
(Reproduced by courtesy of the C. H. Stoelting Co.)

until he succeeds in traversing the maze several times in succession without error. Besides error scores, repetitions and time scores are usually kept.

Carr (12) has devised a stylus maze containing invisible stops so that the subject cannot distinguish the pathways from blind alleys by visual means, and in which, therefore, it is unnecessary to hide the maze from sight.

An interesting type of maze, the high-relief finger maze, has been devised by Miles (42). In this maze the pattern to be traced by the subject is constructed of wire fastened to a board. The subject runs his finger along the wire instead of drawing the stylus through grooves or slots. A description of stylus mazes has been given by Knotts and Miles (36).

## 2. Mirror Drawing

The mirror-drawing test has been much used in the study of trial-and-error learning. In this test (see Figure 22) the subject is required to trace a figure, *e.g.*, a star or a circle, which is visible to him only in the mirror. The reversal of his ordinary habits reduces the subject to trial-and-error learning, as there can be little ideational control. Learning is measured by the number of trials necessary before the subject is able to keep his pencil upon the outline of the figure to be traced, or, if the double line star is used, to keep his pencil between the guide lines. Mirror drawing has been used by Snoddy (65) and by Starch (68) in studying the acquisition of motor habits and trial-and-error learning. Snoddy constructed a double star of brass which was to be traced with a metal stylus. The star outline was so wired

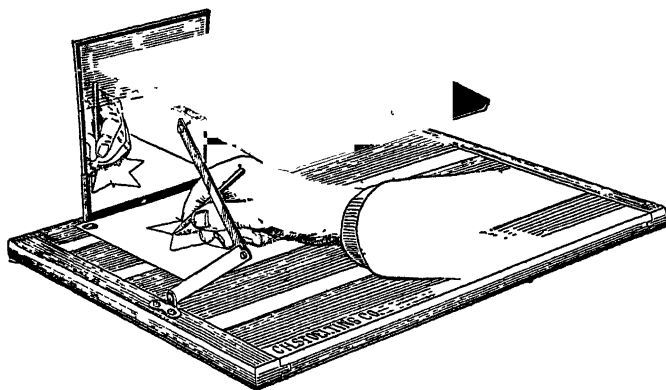


Figure 22.—MIRROR DRAWING TEST (WHIPPLE)  
(Reproduced by courtesy of the C. H. Stoelting Co.)

that any contact of the stylus as it moved along between the double lines was recorded on an electric counter. Burt (8), Calfee (9) and others have used the mirror-drawing test as a measure of adaptability and of learning, and have studied its relationship to general intelligence and to other factors.

## 3. Card Sorting

The card-sorting test was described rather fully in Chapter III, p. 75, as a test of perception. Its possibilities, however, as a measure of the effects of practice, the interference of conflicting habits, and the speed of motor learning, were also stressed. The subject's task is to sort a pack of cards into pigeonholes which bear markings,

*i.e.*, numbers, colors, *etc.*, corresponding to the markings on the cards themselves. Sorting arrangements may be changed from time to time if the acquisition of new habits and the interference of old are to be studied. The table given on p. 79, Chapter III, shows clearly the effect of practice in speeding up the learning time on this test.

#### 4. Results Obtained with Sensory-Motor Tests

(a) *Relationship to General Intelligence Tests.*—In general, the relation of simple sensory-motor tests to measures of verbal or abstract intelligence is fairly low. Warden (74), who administered a stylus maze to a group of forty college students, computed correlations between speed of learning in the maze and scores in the Army Alpha and Thorndike Intelligence Tests. In this group, the correlation of Alpha with the stylus maze was .37; and the correlation of the Thorndike Intelligence Test with the stylus maze was .30. Peterson and Allison (54) tested three groups of college students, fifteen in each group, with a stylus maze. The experimental conditions were different for the three groups. Group 1 was blindfolded throughout the experiment. Group 2 was allowed to inspect the maze ten seconds and then blindfolded. Group 3 was allowed twenty seconds for inspection of the maze before being blindfolded. The correlations of the scores in maze learning (number of repetitions, number of errors and time, combined with equal weights) with the Army Alpha examination were .17 for Group 1, .23 for Group 2, and .02 for Group 3. This is an interesting result. The relation of maze tracing to Alpha is negligible in all three cases. However, since inspection of the maze for a fairly long period (Group 3) reduced the  $r$  to zero, there is a suggestion that the novelty of the task may have been the only reason why any correlation with Alpha appeared. Spence and Townsend (67), using the high-relief finger maze, have studied the relation of maze learning to intelligence test scores in a group of twenty college men. The subjects were divided into two groups, ten with high scores in the Thurstone Intelligence Test and ten with low scores. Correlations of trials, errors and time with the intelligence test scores ranged from — .09 to .52, all of which coefficients are unreliable because of the small size of the sample.

Burt (8) has reported a correlation of .67 between teachers' estimates of intelligence and mirror drawing in a group of thirty elementary-school boys. This high relationship has not been substantiated, however, by other investigators. Calfee (9), for example, obtained



correlations between mirror drawing and school grades of .07 in a group of thirty elementary-school boys, — .07 in a group of fifty-one college men, and .19 in a group of fifty-two college women. Clinton (13), in an extensive study of mirror-drawing ability, computed correlations between ability in tracing a star pattern and measures of general intelligence in groups of elementary, high-school and college students. These groups varied in size from twenty-six to eighty-seven; and the correlations ranged from —.38 to .27.

As already reported on pp. 76, 77, the card-sorting test has little relation to measures of general intelligence. Dewey, Child and Ruml's norms cited on p. 79 exhibit a steady increase in card-sorting scores from nine to thirteen years of age.

(b) *Relationship of Sensory-motor Tests to Tests of Other Abilities*.—Calfee (9) has reported a series of correlations obtained between mirror drawing on the one hand, and card dealing, card sorting and alphabet sorting on the other. The correlations, which ranged from .06 to .37 in the groups of boys and college students described above, seem to indicate that these abilities are fairly specific. Pyle (58) has obtained a correlation of .53 between substitution learning and card sorting in a group of ninety-three college students. This fairly high correlation probably results from the fact that both substitution and card sorting require quick perception and speed of movement, as well as ability to form rather simple associations. Starch (68) has shown that practice produces fairly rapid improvement in mirror-drawing ability at first. This improvement becomes less and less as time goes on, but continues over a long period of time. Weidensall (75), who used a mirror-drawing test in a study of delinquent and criminal women, found the delinquent women to be slower and more variable than the normal, and to make more errors. This author regards the test as a good measure of patience and emotional control.

In the Minnesota Study (49) the correlation of digit-symbol learning with card sorting was .45 (corrected for attenuation) in a group of 217 boys; while the relation of card sorting to shop success was .27 in the same group. There is general agreement (Pyle [59], Calfee [9] and Clinton [13]) that boys are inferior to girls in card sorting; but in mirror drawing the boys are superior up to about age twelve. This last rather queer result is due probably to selection.

(c) *Reliability of Sensory-motor Tests*.—The Minnesota investi-

gators (49) obtained a reliability coefficient of .72 for their card-sorting test. Husband (32), in a group of twenty college students, obtained a reliability for the high-relief finger maze of .95. Scores in the maze were in terms of errors made. Nyswander (47) has studied the reliability of both stylus and high-relief finger mazes using different methods of scoring results. Reliability coefficients varied considerably according to the method used, being highest (e.g., around .79) when scores on the odd-numbered trials were correlated against scores on the even-numbered trials, and the reliability of the whole maze estimated by the Spearman-Brown prophecy formula (20). Yoakum and Calfee (85) have reported a reliability coefficient of .79 for the mirror-drawing test, when the first trial was correlated against the second. This fairly high degree of correspondence is considerably reduced by practice, the correlation between trial 1 and trial 6 being .59. Practice in mirror drawing tends to decrease individual differences, which may explain this drop in correlation.

### Test 2. Mirror Drawing

*Material:* Mirror-drawing outfit; single-line or double-line star blanks; thumb tacks; stop-watch. Mirror-drawing apparatus and blanks may be purchased from the C. H. Stoelting Company, Chicago, Illinois.

*Method:* Place the star blank under the screen so that S can see it only in the mirror. Fasten the blank down with thumbtacks in such a posi-

TABLE XVI  
TIME IN SECONDS FOR MIRROR-DRAWING

Group I comprised thirty elementary-school boys, Group II, fifty-two women, and Group III, fifty-one men in the freshman class of the University of Texas. The averages for each group in each trial are not here reproduced.

(Yoakum and Calfee, from Whipple [77])

Group	Trial	I	II	III	IV	V	VI	Ave.
I	Median	243 0	121 0	93 0	82 0	68 0	50 0	110 33
	M.V.	94 9	45 5	28 1	34 7	24 7	17 1	36 57
	Slowest	517 0	245 0	205 0	180 0	158 0	113 0	210.00
	Fastest	69 0	51 0	41 0	43 0	40 0	32 0	53.66
II	Median	92 0	65 0	48 0	41 0	35 0	28 0	54 70
	M.V.	64 1	33 9	26 6	19 3	21 9	14 2	27 40
	Slowest	700 5	337 5	303 5	153 5	201 8	171 0	242 37
	Fastest	31 5	23 5	19 3	18 3	17 8	17 0	23 95
III	Median	167 5	105 0	80 0	68 0	56 0	48 0	97 83
	M.V.	104 2	39 3	30 3	19 7	19.9	13 5	33 38
	Slowest	752 0	277 0	270 0	175 0	121 0	105 0	193 33
	Fastest	72 0	49 0	40 0	34 0	33 0	23 0	46.87

tion that the starting point indicated by the cross line is on the side farthest away from S. Instruct S to place the point of his pencil upon the cross line and to trace the outline of the star, moving his pencil from left to right. Direct S to work as rapidly and as accurately as possible. If the single-line star is used, S's task is to keep on the line; if the double-line star is used, S's task is to keep within the two lines. Give six trials, using a different blank each time.

*Record:* Keep a record of the time and the number of errors in each trial. For the single-line star, an error is counted each time the subject leaves the line; for the double-line star an error is recorded each time the subject crosses one of the guide lines or touches one of them. Average the time and errors for the six trials; or take the time and errors of the sixth trial as the subject's score.

*Norms:* Norms for the mirror drawing test, in terms of time, are given in Table XVI.

### TESTS OF ASSOCIATION

Tests of association are usually classified under two heads: those which allow almost complete freedom in the subject's responses (free association), and those which place certain restrictions upon the character of the subject's responses (controlled association). Free association may be *continuous*, in which, for example, the subject is asked to give all the words he can think of in three minutes; or *discrete*, in which the subject is asked to respond—with the first word that comes to mind—to each of a list of words presented by the experimenter. In controlled association, restrictions of various sorts are placed upon the subject's responses. He may be told, for instance, to give the opposites of the stimulus words presented; or a "whole" of which the stimulus word is a part; or some other condition may be imposed. Association tests have been employed to study (a) "repressed ideas" (psycho-analytic), (b) personality differences, (c) the associations of the insane, (d) the effects of environmental influences, (e) sex differences, and for many other purposes.

#### 1. Free Association

(a) *Continuous Method.*—A test of free association by the continuous method was placed by Binet at Year XII in his 1911 scale. In this test the child is asked to name as many words as he can in three minutes. A total of sixty words is considered to be a ten-year performance. Terman (71), in his revision of the Binet Tests, placed this test at Year X.

According to Terman, tests of free association by the continuous

method measure not only the readiness of voluntary verbal responses but the richness and variety of previous associations as well. Terman's data (71) show that only 21 per cent. of the eight-year-old children whose records were used in the standardization of the Stanford-Binet Scale were able to pass this test, *i.e.*, name sixty words in three minutes; by age twelve, however, 85 per cent. of the children were able to pass the test. Binet (70) states that young children exhaust an idea by naming it; for example, if the association "hat" is given, children pass on "to another word without noticing that hats differ in color, in form, have various parts, different uses and accessories, and that in enumerating all these, they could find a large number of words." In dull and low-grade children there is often extreme poverty of association, with frequent repetition and much hesitation. Abstract words are rarely given except by bright children.

When S writes, rather than speaks, his responses, the total number of words is, of course, less for a given time. In Table XVII are given the results of Pyle's studies (57), in which the number of words written in three minutes by normal children have been classified by age groups.

TABLE XVII  
NUMBER OF WORDS WRITTEN IN THREE MINUTES—FREE ASSOCIATION (CONTINUOUS METHOD)  
(From Pyle [57])

Boys												
Age .	8	9	10	11	12	13	14	15	16	17	18	Ad.
No.	33	60	66	66	77	80	57	38	36	16	21	64
Norm	23.0	26.9	29.7	33.3	34.2	33.9	33.3	40.0	33.3	42.8	48.9	42.2
Av. Dev.	7.5	7.6	9.0	11.4	10.9	14.6	13.2	14.8	14.6	12.3	16.6	13.8
Girls												
Age. . .	8	9	10	11	12	13	14	15	16	17	18	Ad.
No. . . .	37	82	88	65	90	66	61	46	46	38	29	86
Norm . .	23.7	31.0	32.2	36.8	36.6	38.3	39.1	40.2	40.9	41.6	47.1	38.3
Av. Dev. .	8.2	8.9	10.8	12.1	15.4	16.8	12.9	13.8	14.1	14.0	13.9	13.1

It will be noted that the average number of associations for boys, as well as for girls, increases regularly from the eighth to the eighteenth year. Among the tests which she gave to 200 women college students, Carothers (11) included a word-naming test, in which

the task was to write as many words as possible in three minutes. Carothers' results showed that the average number of words written was 67.50 with a P.E. of 7.50.

Jastrow (33) has made an interesting study of the degree of community shown in the free associations given by men and women. He found that words like book, table, man, *etc.*, were given over and over again, showing the tendency for ordinary thinking, like ordinary activity, to deal with fairly commonplace things. Manchester (40), who also compared the responses of men and women, reported that men's associations tend, on the whole, to be more objective and less personal than women's associations. This early work by Jastrow and Manchester has been repeated by Nöh and Guilford (45) with essentially the same result. The sex differences in association found in the later study, however, were smaller than those previously reported. This leads these authors to suggest that men and women may be becoming more alike in their interests.

(b) *Discrete Method*.—The most detailed experimental studies of free association by the discrete method have been made by Kent and Rosanoff (35). These authors compiled a list of 100 common words which are shown in Figure 23. These words are read in order by the experimenter, and the subject is instructed to reply with the first word that comes to mind. S's reply is written down by the experimenter in the space provided on the test blank. Frequency tables have been drawn up by Kent and Rosanoff (35), based upon the responses to their 100 words given by 1,000 normal adults. From these tables it is possible to compare the responses of an individual with those given by Kent and Rosanoff's standard group. The first word, for example, in the Kent-Rosanoff list is "table," to which, from the frequency tables, we find that 267 normal adults gave the response "chair." If an individual responds "chair" to the stimulus word "table," therefore, his association has a frequency value of 267. Kent and Rosanoff classified the responses given by their 1,000 normal subjects as *common*, *i.e.*, associations which occur in the tables; *doubtful*, *i.e.*, associations which are grammatical variants of responses listed in the tables, *e.g.*, "table-inky" is doubtful since only "table-ink" is found in the list for that stimulus word; and *individual*, *i.e.*, associations which are not found in the tables, and hence were not given by any of the 1,000 normal subjects. Common responses are usually divided into *specific* and *non-specific*. Thus

1. Table	26. Wish	51. Stem	76. Bitter
2. Dark	27. River	52. Lamp	77. Hammer
3. Music	28. White	53. Dream	78. Thirsty
4. Sickness	29. Beautiful	54. Yellow	79. City
5. Man	30. Window	55. Bread	80. Square
6. Deep	31. Rough	56. Justice	81. Butter
7. Soft	32. Citizen	57. Boy	82. Doctor
8. Eating	33. Foot	58. Light	83. Loud
9. Mountain	34. Spider	59. Health	84. Thief
10. House	35. Needle	60. Bible	85. Lion
11. Black	36. Red	61. Memory	86. Joy
12. Mutton	37. Sleep	62. Sheep	87. Bed
13. Comfort	38. Anger	63. Bath	88. Heavy
14. Hand	39. Carpet	64. Cottage	89. Tobacco
15. Short	40. Girl	65. Swift	90. Baby
16. Fruit	41. High	66. Blue	91. Moon
17. Butterfly	42. Working	67. Hangry	92. Scissors
18. Smooth	43. Sour	68. Priest	93. Quiet
19. Command	44. Earth	69. Occan	94. Green
20. Chair	45. Trouble	70. Head	95. Salt
21. Sweet	46. Soldier	71. Stove	96. Street
22. Whistle	47. Cabbage	72. Long	97. King
23. Woman	48. Hard	73. Religion	98. Cheese
24. Cold	49. Eagle	74. Whiskey	99. Blossom
25. Slow	50. Stomach	75. Child	100. Afraid

Figure 23.—FREE ASSOCIATION TEST (KENT-ROSANOFF)

"hammer-nail" is a specific response, "hammer-large" non-specific. Still another classification suggested by Kent and Rosanoff is that of *normal* responses. It is clear that the responses gathered from even 1,000 normal persons cannot possibly cover all of the legitimate replies; and hence Kent and Rosanoff (Appendix to their frequency tables [60]) have given certain rules designed to indicate what responses are to be considered "normal."

It is interesting to compare the individual responses of Kent and Rosanoff's 1,000 normal subjects with the individual responses given by 247 insane patients.

TABLE XVIII  
RESPONSES OF VARIOUS GROUPS TO KENT-ROSANOFF FREE ASSOCIATION TEST (60)

Subjects	Common Reactions		Doubtful Reactions %	Individual Reactions %	Failures of Reaction %
	Specific %	Non-specific %			
1,000 normal adults	85.5	6.2	1.5	6.8	
247 insane adults	66.4	4.3	2.5	26.8	
253 defective children aged over 9 years	75.2	8.2	2.1	13.0	1.5
125 normal white children, 11-15 years	82.0	7.2	1.6	8.6	0.6
175 normal white children, 4-10 years	62.7	4.2	3.2	18.8	11.1
125 normal Negro children, 11-15 years	75.3	7.2	2.5	14.9	0.1
175 normal Negro children, 4-10 years	54.1	3.5	2.5	33.2	6.7

Only about 7 per cent. of the responses given by the normal group were classified as individual, while 27 per cent. of the responses given by the insane were so classified. This result demonstrates experimentally the existence of those peculiarities and eccentricities which are often observed in the thinking of the insane. It will be noted that young Negro children also gave a large percentage of individual replies. It may be that such peculiarities of association are direct outgrowths of the social restrictions and deprivations which generally surround this group. There appear to be significant qualitative differences between the responses given by the insane and the normal. Kent and Rosanoff reported that the insane give more incoherent and stereotyped replies; also that they frequently give responses which rhyme with the stimulus word. Murphy (44) has made an extensive study of the Kent-Rosanoff Test in groups of 250 normal, 120 dementia præcox and eighty-two manic-depressive individuals. Responses were carefully classified into thirteen cate-

gories, such as contiguity, similarity, contrast, adjective-noun, *etc.* Murphy found that the manic-depressive group tended to give more rhyme and sound associations than did the dementia præcox group; but no marked differences appeared between these groups in type of association.

One of the earliest uses of the free association test as a means of diagnosing emotional difficulties or "complexes" was made by Carl Jung (34). Jung's method was to present to his subject, or patient, a set of stimulus words covering a wide range of topics. Some of these words were selected so as to have emotional value for the subject. These so-called "critical" words were mixed in with the indifferent, or innocuous words. Critical words as used by Jung are illustrated by the following: dead, sick, to pray, to fear, anxiety, to kiss, bride, contented. The theory underlying the free association method as used by Jung and by other psycho-analysts is that extreme embarrassment, timidity, useless fears, worries, *etc.*, when they occur in highly nervous or neurotic people, generally center around forgotten or little-understood emotional episodes in the person's life. Those words in the list which "remind" the subject of these occurrences will tend to provoke personal or highly emotional associations, often accompanied by laughter or by blushes. Long reaction time to a given stimulus word, repetition of the stimulus word, or a refusal to respond at all, are interpreted by the analyst as constituting an avoidance of unpleasant associations called up by the stimulus word. Such significant reactions are called "complex indicators"; and it is these responses that are followed up by the analyst in an effort to relieve the patient's difficulties.

The time of response to the single items in a free association test may be taken with an ordinary one-fifth-second stop-watch. Kent and Rosanoff did not record the time of responses. Lengthened responses, however, may be important as indicators of emotional states, as has been pointed out by Smith (64), and by Landis, Gullette, and Jacobson (37).

The "detective" use of the free association test should be mentioned in connection with the use of free association in ferreting out emotional difficulties. The "detective" method is based upon the assumption that even in presumably voluntary associations a subject's reply will be inevitably tied up with those experiences suggested by the stimulus word. So, if an individual has committed a crime—a



theft, say—words bearing upon the circumstances of the crime will elicit responses which will serve to “give the subject away.” If the subject tries to “beat the game” by giving some foolish or irrelevant association, the time of responses will ordinarily be lengthened. The detective use of free association has distinct possibilities which have not up to the present been fully realized. This method was first employed by Münsterberg (43), and has given some striking results (38).

Several studies of the free association test have been made with children. Rosanoff and Rosanoff (61) have reported that children exhibit a decided tendency to repeat a previous response; and that they also give more individual responses than do adults. Woodrow and Lowell (81), however, who have compiled frequency tables especially for children, found that children give fewer individual associations than adults, when their associations are evaluated in terms of their own frequency tables. Otis (48), in a study of 130 feeble-minded and 200 normal children, found that the feeble-minded give an unusual number of non-specific (*i.e.*, general) and phrase responses, show a greater tendency to repeat the response word, and often fail to respond at all. Wheat (76) has devised a free association test consisting of twenty-five words picked at random from the 500 most commonly used English words. This test was given to 1,323 children in grades 4 to 8. The reliability of the test, when given a second time after a week's interval to 111 sixth-grade children, was .74.

The relationship of free association to general ability will depend largely upon the measure of free association employed. Conrad and Harris (15) obtained correlations which ranged from — .52 to .74 between N.I.T.<sup>1</sup> scores and various classifications of free association responses in a group of 166 children, eleven to fifteen years old. Probably the most interesting finding in this study is the correlation of — .53 between N.I.T. scores and the total number of “failures to respond.”

### Test 3. Free Association (Discrete Method)

*Material:* Test blanks containing the 100 stimulus words used by Kent and Rosanoff (see Figure 23); the Kent-Rosanoff frequency tables; stopwatch. Test blanks may be purchased from the C. H. Stoelting Company, Chicago, Illinois.

<sup>1</sup> National Intelligence Test.

*Method:* This test should be given in a room as quiet and as free from distracting influences as possible. E should instruct S as follows: "I am going to read to you one at a time a series of 100 words. Before each word I shall give the signal 'Ready.' As soon as you hear the stimulus word respond with the first word that comes to mind. Do not repeat the word which I have spoken, and do not respond with a phrase. Your response must be a single word, and you must answer as quickly as possible." E writes down the response given by S in the space provided on the blank.

*Variation in Method:* The time taken by S to respond to each of the stimulus words may be recorded with a stop-watch. E should start the stop-watch as soon as he has spoken the stimulus word, and stop it as soon as S has spoken or has written his response.

*Record:* In order to determine the degree to which S's responses correspond with those given by other people, the frequency of each response should be found from the Kent-Rosanoff frequency tables. If S's response has a frequency value of 10 or 50, this number should be written on the blank opposite the stimulus word. If S's response is "individual," that is, not found in the frequency tables, its frequency value is zero. (The meaning of the classifications, *common*, *normal*, *doubtful* and *individual* will be found on p. 103.) In order to find the median frequency of S's responses, arrange the separate frequency values of the 100 words in order of size. Count down until the fiftieth value is reached; the point midway between the fiftieth and the fifty-first frequency values is the median frequency. The median frequency is an index of the extent to which S's responses agree with those of other people. A high median frequency may mean that S's thinking is ordinary and commonplace; a low median frequency that it is individual and different. S's responses may also be classified in the various categories given by Kent and Rosanoff. The percentage of *common*, *individual*, *doubtful* and *failure* responses may also be computed.

If the time of S's separate responses has been recorded with a stop-watch, these may be put into a frequency distribution, and S's median response time compared with that of other members of his group.

*Norms:* The Kent-Rosanoff frequency tables may be found in the following references (35, 60). Table XVIII gives the results found by Kent and Rosanoff for different groups of subjects.

## 2. Controlled Association

Simple tests of controlled association measure largely the speed and facility with which certain familiar associations can be reinstated. The subject's response is so restricted by the instructions that ordinarily only one reply is possible. For example, in an opposites test there is only one reply to the stimulus word "high." Controlled

association tests as they become more difficult involve to a greater and greater degree the ability to "reason out" or educe relations, which are often highly abstract. Spearman (66) holds that the more difficult controlled association tests, as for example, opposites or analogies,<sup>1</sup> are among the best tests of general intellectual ability. Opposites and analogies tests are found in many of the standard general intelligence examinations, such as Army Alpha, and the Terman and Otis group tests.

The first systematic study of controlled association tests was made by Woodworth and Wells (82). These authors devised tests which involved such relations as opposites, part-whole, whole-part, agent-action (noun-verb), adjective-noun, coördinates and others. The method used by Woodworth and Wells was to present to the subject a list of twenty words. The subject was instructed to respond orally to each stimulus word and the time for all twenty responses was taken by the experimenter. The average time of response of adults to a simple opposites test is given as 1.11 seconds by Woodworth and Wells. The average time of response of the same subjects to the part-whole test was 1.53 seconds; and the average time for the mixed relations (or analogies) test was 3.14 seconds. As might be expected, the mixed relations test is the most difficult type of controlled association as measured by time of response, while simple opposites is the easiest. The opposites, part-whole and mixed relations tests with which Woodworth and Wells obtained the results quoted are shown in Table XIX.

Various methods have been used in administering controlled association tests. The words may be shown one at a time, the subject's response written down and his reaction time taken. Or the test may be given by the group method, in which case the subjects write down their own responses. When only the total time taken by the subjects in answering all of the words in the test is recorded, the average time of association per word is found by dividing the total time by the number of words in the list.

Controlled association tests have in general shown high correla-

<sup>1</sup> In analogies tests the instructions are to discover the relationship between the first and second words, and then fill in the blank space with a word which bears the same relationship to the third word.

e.g., eye—see::ear—?  
wing—bird::fin—?

TABLE XIX  
CONTROLLED ASSOCIATION LISTS  
(From Woodworth and Wells [82])

Opposites Test	Part-whole Test	Mixed Relations Test	
high	elbow	Eye—see	Ear—
summer	hinge	Monday—Tuesday	April—
out	page	Do—did	See—
white	finger	Bird—sing	Dog—
slow	wing	Hour—minute	Minute—
yes	morning	Straw—hat	Leather—
above	blade	Cloud—rain	Sun—
north	mattress	Hammer—tool	Dictionary—
top	chimney	Uncle—aunt	Brother—
wet	cent	Dog—puppy	Cat—
good	sleeve	Little—less	Much—
rich	brick	Wash—face	Sweep—
up	deck	House—room	Book—
front	France	Sky—blue	Grass—
long	pint	Swim—water	Fly—
hot	fin	Once—one	Twice—
east	steeple	Cat—fur	Bird—
day	month	Pan—tin	Table—
big	hub	Buy—sell	Come—
love	chin	Oyster—shell	Banana—

tions with tests designed to measure general intelligence. Spearman (66) has reported a correlation of .89 between the opposites test and measures of "general mental ability." In Chapter I, p. 44, the average correlation of the opposites test (synonym-antonym) with the other tests of the Army Alpha will be found to be .77. Bonser (6) reported a correlation of .85 between an opposites test of twenty items and the average of three "reasoning" tests. Wyatt (83) found a correlation of .67 between the part-whole test and teachers' estimates of intelligence in a group of thirty-four children, eleven to thirteen years old; and correlations of .62 and .80 between an analogies test and measures of intelligence. Wylie (84) studied the relation of opposites tests to general intelligence in groups of from 73 to 149 school children, ten to fourteen years of age. The correlations ranged from .65 to .77.

The intercorrelations of controlled association tests are generally high. Carothers (11), for example, found a correlation of .57 between the Woodworth-Wells opposites test and the Woodworth-Wells mixed relations; while Schneck (63) obtained a correlation of .72 between analogies and opposites in a group of 210 college students.

Pyle obtained the results shown in Table XX when the opposites of the following twenty words were written by the subjects:

## OPPOSITES

(From Pyle [57])

1. good	11 like
2 outside	12 rich
3. quick	13 sick
4. tall	14. glad
5 big	15. thin
6 loud	16 empty
7 white	17. war
8. light	18. many
9. happy	19. above
10. false	20. friend

TABLE XX

NUMBER OF CORRECT ASSOCIATIONS WRITTEN IN SIXTY SECONDS OPPOSITES TEST

(From Pyle [57])

## Boys

Age	8	9	10	11	12	13	14	15	16	17	18	Ad.
No.	33	65	60	61	72	65	61	40	33	17	22	62
Norm.	9.0	8.4	7.5	10.9	11.5	14.5	14.5	16.0	18.6	19.6	22.4	22.1
Av. Dev.	3.3	3.0	3.1	2.9	2.9	4.5	4.3	5.2	5.3	3.3	3.2	3.3

## Girls

Age	8	9	10	11	12	13	14	15	16	17	18	Ad.
No.	33	56	77	65	74	73	58	49	48	27	26	85
Norm.	8.0	7.6	10.9	11.2	13.9	14.9	17.4	17.3	19.3	21.4	23.4	23.4
Av. Dev.	4.0	2.9	3.1	3.0	3.6	4.3	3.9	5.1	4.2	4.9	3.1	4.0

It appears from the table that girls are somewhat faster than boys in the opposites test, although the differences are slight.

Hollingworth (30) has made an interesting study of the effects of practice upon speed in the opposites test. Eleven men and eight women were given 100 trials with an opposites test of fifty words. Their speed in reading aloud the opposites to the fifty words from a typewritten sheet, which contained the stimulus words and their opposites, was also measured. It was found that intensive practice had greatly increased the speed of response in the test proper; but that even after ninety-five trials there was considerable time taken in the test itself which could not be accounted for by the silent reading of the stimulus words, and the articulation of their opposites. It seems clear, therefore, that no matter how easy the test a certain amount of mental effort is demanded.

The reliability of controlled association tests is high. Wylie (84)

has obtained reliabilities of .80 to .95 for an opposites test of thirty-five items. Schneck (63) obtained a reliability of .93 for a difficult opposites test containing 130 items; and a reliability of .88 for a difficult analogies test containing forty items.

### 3. Other Controlled Association Tests

There are several other controlled association tests which deserve to be mentioned. The most important of these is the vocabulary test, which is described in Chapter I, p. 28, as being probably the best single test of abstract or verbal intelligence. Terman (70) found that this test gives an exceedingly close approximation to the mental age obtained from the Stanford Revision of the Binet Tests. The vocabulary test included in the Stanford-Binet Scale was arranged and standardized by Terman and Childs (73). Their method was to select the last word of every sixth column in a dictionary containing the 18,000 most common English words. Terman's list of 100 words is based on the assumption that the words selected according to this arbitrary rule will furnish a sample which will yield a fairly reliable index of the subject's vocabulary. This test is administered as an individual test, the subject's definitions being given orally and written in by the experimenter. An individual's total vocabulary is equal to his total score multiplied by 180. Table XXI gives the vocabulary scores for different mental ages computed by Terman.

TABLE XXI  
MENTAL AGE NORMS FOR BOYS AND GIRLS,  
ON THE TERMAN VOCABULARY TEST  
(From Terman [72])

Mental Age	Boys	Girls
6-6 to 7-5 . . . . .	13 6	12 5
7-6 to 8-5 . . . . .	20 7	17
8-6 to 9-5 . . . . .	24	21 8
9-6 to 10-5 . . . . .	31	31
10-6 to 11-5 . . . . .	35 7	32 5
11-6 to 12-5 . . . . .	42 5	40 5
12-6 to 13-5 . . . . .	45	49
13-6 to 14-5 . . . . .	50 5	51
14-6 to 15-5 . . . . .	51 7	56 5
15-6 to 16-5 . . . . .	61 6	61 8
16-6 to 17-5 . . . . .	65	68
17-6 to 18-5 . . . . .	73	70
18-6 to 19-5 . . . . .	75	76 2

Probably the most satisfactory method of giving the vocabulary test by the group method is to present a list of words for each of which four or five possible synonyms are supplied (method of multi-

ple choice). The subject is instructed to underline that word which best defines the stimulus word. Thorndike has compiled several vocabulary lists for grades 3 to 10, which may be purchased from the Bureau of Publications, Teachers College, Columbia University. For advanced adult groups, the vocabulary tests in Thorndike's CAVD (Chap. I, p. 36) will be useful.

The form-naming test and the color-naming test may perhaps be best classified as controlled association tests. These tests were devised by Woodworth and Wells (82). In the form-naming test the subject is required to name orally as rapidly as possible 100 geometrical figures printed on a sheet. The figures are a circle, square, cross, triangle and star repeated over and over in random order. In the color-naming test the subject names as rapidly as he can 100 small squares of color. These color squares are red, blue, yellow, green and black repeated over and over in random arrangement. Both of these tests have been used as measures of perception and attention as well as of association. Women are generally faster than men in both form and color naming. Whitley (78) found that women take on the average sixty-seven seconds to name the 100 colors, while men take eighty-five seconds. The correlation between form naming and color naming was reported by Whitley to be .73.

#### Test 4. Controlled Association

*Materials:* Test blanks containing opposites, part-whole, whole-part, mixed relations, and other blanks from the Woodworth-Wells series; stop-watch. Controlled association blanks may be purchased from the C. H. Stoelting Company, Chicago, Illinois.

*Method:* Place the test blank, say, the opposites, face down before S. Instruct S to turn over the blank at the word *go* and to give orally the opposite of each of the twenty words on the blank in succession. If other controlled association blanks are used, several illustrations of what is wanted should be given before the test is administered. Be sure that S understands what he is to do.

*Alternate Method:* Controlled association tests may be given by the group method, the responses being written instead of spoken. Distribute the test blanks, face down, to the subjects. Explain and illustrate what kind of response is required; then at the word *go* tell the subjects to turn over their blanks and write in the correct associates as quickly as possible. If the work-limit method is used each subject must read the time required to complete from a time clock. If the time-limit method is used thirty seconds or even twenty seconds is probably long enough for adult subjects.

*Record:* With adult subjects, usually only the time score need be con-

sidered, since errors are rarely made. In the case of children, an S upon giving a wrong response, may be stopped and required to give a right one before going on: or a penalty may be added to the time score. The average time per association is obtained by dividing the total time taken on the blank by 20.

*Norms:* Norms obtained by Woodworth and Wells (adult subjects, oral response) for the opposites, part-whole, and mixed relations tests will be found on p. 114. Pyle's norms for his list of opposites (p. 116) are given in Table XX.

## TESTS OF MEMORY

The phenomena of memory may be classified under the heads of fixation, retention, recall and recognition. Fixation, or the ability to get new impressions quickly and accurately, is measured by tests of immediate memory. Retention, or retentivity, is measured by tests of recall and recognition, which may be given at varying time intervals after the original presentation of the memory material. Tests of memory, themselves, may be classified as tests of "rote" memory, or memory for disconnected impressions; and tests of "logical" memory, or memory for connected meaningful material. In rote memory tests the subject is required to reproduce the material exactly as it was presented, *i.e.*, he must learn it "by heart." In logical memory tests the ideas or meanings involved are asked for, rather than an exact reproduction of the material as presented.

### I. ROTE MEMORY

#### 1. Simple Immediate Memory

(a) *Memory-span Method.*—Memory span is the longest series of items, *e.g.*, letters, digits, words, *etc.*, which the subject is able to reproduce correctly after a single presentation. Digits are the most frequently used material in memory-span tests. Presentation may be visual or auditory. In testing memory span for digits, by the visual method, the material is presented upon a series of cards which contain digits printed in large type. These test cards contain from four to twelve digits each (fifteen may be taken as the upper limit if college students are subjects). E begins with the first card, the one containing four digits, and shows it to S for *four* seconds. As soon as the card is covered, S writes down or gives orally all of the digits which he can remember. S is instructed to give the digits in the order in which they were presented. Each card is shown for as many



seconds as there are digits printed upon it. Time of exposure is controlled by a metronome. In auditory presentation, the digits or letters are read at a rate of one per second. Subjects are cautioned to wait until the presentation is complete, before writing down, or giving orally, items which they have heard.

The length of the memory span increases during the early years. Hallowell (27), in testing by the auditory method the digit span of 413 children ranging from twelve to forty-seven months in age, found that "not until forty-six months or almost the fourth birthday do at least half of the children have a span of 4 or more." Hurlock and Newmark (31) have reported an average auditory memory span of five digits for children ranging in age from four years nine months to five years ten months. In the Stanford-Binet an auditory memory-span test of three digits is placed at Year 3; four digits at Year 4; five digits at Year 7; six digits at Year 10; seven digits at Year 14; and eight digits at the superior adult level. Terman's method is to give the child three trials, recording the test as passed if one list out of the three is reproduced correctly. Pyle (57) has published extensive data on auditory memory span for concrete and abstract words. His method of administering this test is to read to the subjects lists of from three to eight words. Each word remembered correctly and in its correct position in the list counts two points; one point for reproduction and one point for correct position. Since there are thirty-three words in all in Pyle's lists, the maximum score obtainable

TABLE XXII  
MEMORY SPAN (CONCRETE WORDS) SCORED BY METHOD GIVEN ABOVE  
(From Pyle [57])

Boys												
Age... . . .	8	9	10	11	12	13	14	15	16	17	18	Ad.
No. . . . .	31	58	61	55	60	60	35	25	14	7	5	61
Norm . . . .	31.2	32.4	35.8	37.7	37.7	38.3	40.0	40.2	43.4	45.7	49.0	44.3
Av. Dev. . .	6.7	7.4	6.3	6.4	5.0	5.6	6.4	4.9	6.3	5.1	7.6	6.6
Girls												
Age..... . .	8	9	10	11	12	13	14	15	16	17	18	Ad
No. . . . .	37	68	69	52	70	51	34	13	17	8	2	88
Norm . . . .	32.9	32.7	39.6	37.7	38.7	40.1	41.2	42.0	42.5	40.5	52.0	47.6
Av. Dev. . .	7.1	6.2	5.2	5.2	6.1	5.4	7.0	7.0	4.8	4.6	2.0	7.7

is 66. Pyle's results for boys and girls from eight to eighteen years of age are given in Table XXII. Pyle's memory material is as follows:

*Concrete Words*

1. street, ink, lamp
2. spoon, horse, chair, stone
3. ground, clock, boy, chalk, book
4. desk, milk, hand, card, floor, cat
5. ball, cup, glass, hat, fork, pole, cloud
6. coat, girl, house, salt, glove, watch, box, mat

At the adult level the memory span for digits varies from 7 to 9. Carothers (11), in a group of 200 freshmen women, found the average auditory memory span for digits to be  $7.53 \pm .53$ . In an extensive study of memory, Anastasi (2) tested 225 male college students ranging in age from twenty to twenty-nine years. In this group the visual digit span was 9.3, S.D. 1.1. This result agrees with that of Garrett (19), who obtained an average visual digit span of 9.1, and an average auditory digit span of 8.4, in a group of 158 male college students. Visual digit span is nearly always slightly higher than auditory digit span. This result arises no doubt from the fact that grouping is greatly facilitated, and review is possible, in visual presentation.

Another type of memory-span test is that of memory for sentences. In his 1911 scale, Binet placed the test of repeating a sentence of ten syllables (which had been heard once) at Year 5; and the test of repeating a sentence of twenty-six syllables at Year 15. Terman (70) placed tests of sentence memory at the following age levels in Stanford-Binet: six to seven syllables at Year 3; twelve to thirteen syllables at Year 4; sixteen to eighteen at Year 6; twenty to twenty-two at Year 10; twenty-eight at Year 16, *i.e.*, average adult level. Terman counts the sentence-memory test passed when a child can repeat correctly one of the three sentences which he has heard; or two out of three with one error in each. At Year 16, however, one of the two sentences heard must be absolutely correct.

(b) *Relation of Memory Span to Other Functions.*—Like most tests of narrow functions, memory span is more closely related to general ability in children than in adults. Hurlock and Newmark (31) have reported a correlation of .59 between auditory

memory span for digits and Stanford-Binet I.Q. in a group of twenty pre-school children. This correlation is somewhat too high because digit-span tests occur at Years 3 and 4 in the Stanford-Binet. Goodenough (26) has studied the digit-span test in relation to M.A. as found from the Kuhlmann-Binet Scale (p. 10). In a group of 100 three-year-olds, the correlation of success in the three-digit test (bi-serial  $r$ ) and Kuhlmann-Binet M.A. was .67; in a group of 100 four-year-olds, the correlation between success in the five-digit test and Kuhlmann-Binet M.A. was .60. Both Binet and Terman have considered the digit-span test to be a good measure of active attention, and hence indicative of general intelligence in children. Terman's data for the Stanford Revision (71) showed a steady increase in digit span with age from three to fourteen years. Norsworthy (46) found that for related and unrelated words, respectively, only 5 and 6 per cent. of feeble-minded children had a memory span equal to the average normal child of the same age.

In a group of 121 college students Wissler (80) obtained the low correlation of .16 between digit span and college grades. Garrett (19) obtained a correlation of .18 between visual digit span and the Thorndike Intelligence Examination in a group of 158 students; and a correlation of .21 between auditory digit span and intelligence in the same group. Carothers (11) calculated the average correlation of digit span with nineteen other mental tests to be only .17. The highest correlation given by the digit-span test in Carothers' study, *i.e.*, .77, was with a number cancellation test.

There is a substantial relationship between auditory and visual memory span. Garrett, in the study mentioned above, obtained a correlation of .57 between auditory and visual digit span. A correlation of .57 was also obtained by Gates (25) between auditory and visual digit span in a group of 172 boys and girls in the fifth and sixth grades. In a group of 197 college students, Gates (23) obtained a correlation of .25 between auditory digit span and logical memory for a prose selection. Anastasi (2) has reported correlations between visual memory span for digits and seven rote memory tests of recall and recognition which range from — .03 to .25.

The narrow range of scores in memory-span tests, and the low correlations of these tests with other measures in adult groups, indicate that memory span is of little value in the upper age levels.

Practice and training play a large rôle in memory span in older groups. As a measure of attentive observation and concentration, memory span is of value with young children; but its low correlation with verbal tests of general ability suggests that the digit span has been much overemphasized in the upper levels of the Stanford-Binet. Investigators are generally agreed that girls are superior to boys in memory span (57, 77).

(c) *Reliability of Memory-span Tests.*—Tests of memory span have in general shown a fairly high degree of reliability. Carothers (11) reported a reliability of .83 for an auditory digit-span test ( $N=45$ ). Hartmann (28) obtained a reliability of .73 for auditory digit span in a group of sixty-three college students; and Anastasi (2) obtained a reliability of .74 for visual digit span in a group of 225 students.

#### Test 5. Memory Span for Digits

*Material:* Test cards containing from four to twelve (or fifteen) digits; metronome. This material may be made by the experimenter or it may be purchased from the C. H. Stoelting Company, Chicago, Illinois.

*Method:* (a) Auditory presentation. Set the metronome at 60. Begin with the four-digit card and pronounce each digit one after the other in time with the metronome beats. Continue with the 5, 6, 7, *etc.*, digit lists. Instruct S, as soon as a series has been presented, to write down immediately as many of the digits as he can remember in their right order.

(b) Visual presentation. Set the metronome at 60, as before. Show the cards in order allowing one second for each digit, *i.e.*, four seconds for the four-digit card, five seconds for the five-digit card, and so forth. In exposing a card the experimenter should count silently in time with the metronome. Lift the card up on the count 0; then count 1, 2, 3, 4, *etc.*, laying the card down immediately after the last count. Give the cards in order beginning with the four-digit card. Instruct S to write down, at the end of each presentation, as many digits as he can remember.

*Record:* For both auditory and visual presentation S's score is the largest number of digits reproduced without error. More accurate results are obtained if two or more lists are given and the results averaged. One may use Terman's method of allowing three trials for each list of digits, counting the test as passed if one list is reproduced correctly.

*Norms:* Norms for children and adults will be found on p. 121. The following table from Bronner, Healy, *et al.* (7), gives digit-span norms (auditory method) for children, by age. These data were compiled by Bronner, Healy, *et al.*, and by Starr (69).

## PSYCHOLOGICAL TESTS

TABLE XXIII

## AUDITORY DIGIT SPAN BY AGE

Age	Bronner, Healy, <i>et al.</i>	Starr
4		4
5		4
6		5
7	5	5
8	5	5
9	5	5-6
10	5	6
11	6	6
12	6	6
13	6-7	6-7
14	6-7	6-7
15	7	7
16	7	
17+.	7	

## 2. Recall Memory

(a) *Method of Retained Members*.—Either fixation or retention may be measured by the method of retained members, depending upon whether the memory test is given immediately or after a time interval has elapsed. In the method of retained members more items are presented than can possibly be reproduced by the subject. The measure of memory is the amount of material (*i.e.*, number of items) which the subject is able to recall. Various kinds of materials may be employed in recall memory tests: words, nonsense syllables, poetry, proverbs, pictures, geometrical figures, and actual objects. In most cases, presentation may be by either the visual or the auditory method.

(b) *Method of Paired Associates*.—The method of paired associates may also be used for testing immediate or delayed memory. Both auditory and visual presentation is employed. In the visual method a series of cards containing a list of paired terms, *e.g.*, man-sky, desk-stone, *etc.*, are presented to the subject one at a time. Directly after the presentation, which may be repeated several times before the test proper, immediate memory is tested for by exhibiting in succession the first item of every pair which has been shown. The subject's task is to write down, or give orally, the term paired with each stimulus item in the original presentation. The score is the number of paired associates which are correctly reproduced. In auditory presentation the pairs of items are read to S one pair at a time. S is then required to write down the second word of each pair upon hearing the first. In order to prevent the learning of associates in serial order the cards are shown in a different order in the test

proper from that employed in the presentation series. Not only words but nonsense syllables, digits or pictures may be used as memory material; or a word may be paired against a nonsense syllable, or a picture against a number.

(c) *Relation of Recall Memory to General Intelligence.*—Except in the case of children, tests of simple recall memory, like tests of memory span, have shown in general little relationship to measures of intelligence. Achilles (1) found that recall of words, geometrical and other forms, and nonsense syllables increased regularly with age and grade. Carey (10) obtained a correlation of .38 between three word recall tests (both auditory and visual) and teachers' estimates of scholastic intelligence. Carey's subjects were 150 children seven to fourteen years old. Gates (23), in a group of 318 grade-school children, obtained correlations of .43 and .44, respectively, between teachers' estimates of intelligence and immediate and delayed recall of nonsense syllables. McGeoch (41) found that in a series of paired associates tests in which nonsense syllables were matched with words, the retention of a normal group of children was never greater than 39 per cent. of the retention scores made by a gifted group.

Bolton (5), in a group of 200 freshmen college women, obtained a correlation of only .18 between word recall and the Otis Self-Administering Test. The correlation between a paired associates test of names and dates and the Otis Test was .38. Both Carothers (11) and Anastasi (2) found low correlations between word-recall tests and vocabulary (a good measure of general intelligence, p. 28). Anastasi reports a correlation of .06 between a difficult word recall test and vocabulary. Her word recall test consisted of forty four-letter words which were presented visually in two series of twenty each. Carothers obtained a correlation of .14 between the recall of twenty-five words presented visually and a vocabulary test ( $N = 100$  freshmen women).

(d) *Relation of Recall Memory Tests to Each Other and to Other Abilities.*—Tests of recall memory for symbolic material have been shown by Anastasi (2) to possess a common "memory factor." Anastasi constructed four paired associates tests as follows: word-word; picture-number; form-number; color-word. The average inter-correlation (corrected for attenuation) of these four tests of recall memory was .58, and the average corrected correlation of the word-

recall test with these four tests was .57. These results suggest a fairly substantial community of function in tests of this sort. In a later study, however, this memory ability did not prove to extend to other performances. Anastasi (3) obtained an average correlation of .02 between a word-word paired associates test and four other very different tests of memory. These were tests of logical memory for difficult prose; delayed memory for words, the test being given after a forty-eight-hour interval; memory for tones;<sup>1</sup> and memory for movement. The subjects were 170 college students. This finding indicates that memory ability is fairly specific except where the material is closely related.

Carothers (11), in a group of 100 women, obtained an average correlation of .13 between word recall and nineteen other tests such as tapping, cancellation, controlled association, logical memory, *etc.* Garrett (19) obtained correlations ranging from —.05 to .59 between two paired associates tests and six memory and learning tests. The correlation of paired associates (auditory) and paired associates (visual) was .59, as contrasted with the correlation of —.05 between paired associates visual and auditory memory span for digits.

(e) *Reliability of Recall Memory Tests.*—The reliability of recall memory tests depends greatly upon the character of the group studied, number of items in the series, and the difficulty of the material. Bolton (5), who employed a list of twenty words, has reported a reliability for word recall of .47. Anastasi, who combined two word-word paired associates tests, each test containing twenty items, obtained a reliability of .88 for the whole series.

### 3. Recognition Memory

The method of recognition, like the methods of recall memory described above, may be used to study fixation, or retention after an interval of time. In a typical recognition test twenty-five items, words or pictures, say, are shown to the subject. These twenty-five items are then mixed with twenty-five more, and the whole series of fifty again presented to the subject, who is instructed to indicate which items he has seen before. Errors are of two kinds: (1) Failure to recognize an item seen before; (2) false recognition of items which were not previously shown. The subject is often asked to express his degree of confidence in his judgment. The score in a recognition test

<sup>1</sup> This test was given by means of the Seashore record (Chapter IV, p 168)

is usually found by deducting twice the number of errors from the total number of items shown. Many kinds of material have been employed in studying recognition memory: words, nonsense syllables, geometrical forms, pictures, designs, advertisements, etc.

(a) *Relationship of Recognition Memory to Recall Memory and to Other Abilities*.—In studying recognition, Bolton (5) presented to 200 college women a list of twenty-four words which were later combined with seventy-two new words and shown again. The subjects were asked to check those words in the new list which they had seen before. The correlation of this word-recognition test with a syllable-recognition test was .46; and with word recall .23. Achilles (1) administered recognition tests of words, geometrical and other forms, proverbs and nonsense syllables to ninety-six adults and more than 600 children. Her results show proverbs to be the easiest material to recognize, nonsense syllables to be the most difficult. The correlations of Achilles' recall and recognition tests averaged .23 for adults and .21 for children. The highest correlations were between those tests in which the material was closely similar. Differences in the familiarity of the subjects with the material, and differences in the length and difficulty of the tests, probably operated to reduce many of these correlations.

Anastasi (2) has obtained low correlations between recognition tests when the materials employed were dissimilar. The correlation between recognition of forms and recognition of nonsense syllables, for instance, was .10. But the correlation between word recall and nonsense-syllable recognition was .26. Apparently, the material was more important than the method in these memory tests. Both Anastasi and Carothers found low correlations between recognition memory and general intelligence. In both recall and recognition memory, females are consistently superior to males (1).

(b) *Reliability of Recognition Tests*.—Bolton has reported a reliability of .48 for the word-recognition test mentioned above. Carothers (11) obtained a reliability of only .33 for a recognition test of twenty-five words in a group of forty-five college women; and a reliability of .73 for a recognition test of twenty-five proverbs. Anastasi, who employed a recognition test of forty nonsense syllables, obtained a reliability of .68 in her group of 225 college students. Short and easy recognition tests give low correlations because of the



large chance element present; longer and more difficult tests show higher reliabilities.

#### Test 6. Recall and Recognition of Words (Methods of Retained Members and of Recognition)

*Material:* E should prepare two lists of twenty-five disconnected words, (a) and (b); and two lists of fifty words each, the first list containing the twenty-five words in (a), the second list the twenty-five words in (b). Two of these lists are for visual and two for auditory presentation.

*Method:* (1) Visual Presentation.—(a) Recall: Give S's sheets containing the first list of twenty-five words and instruct them to study the list carefully for one minute. At the end of this time collect the lists, and have S's write down as many of the words as they can remember. (b) Recognition: Give S's sheets containing a list of fifty words, twenty-five of which have already been seen in (a). Instruct S's to check off on this list all the words which they saw before.

(2) Auditory Presentation.—(a) Recall: Read a list of twenty-five words slowly, allowing about one second per word with a two-second interval between words. Immediately after presentation, instruct S's to write down all of the words which they can remember. (b) Recognition: Instruct S's to prepare a record sheet containing spaces numbered from 1 to 50. Now read the list of fifty words made up of the twenty-five words already heard, mixed in with twenty-five new ones. As each word is read, direct S to mark Y if he has heard it before, N if he has not. Caution S's to put each response opposite its proper number and to leave the space blank if they are not sure.

*Record:* The score for recall memory is the number of words remembered, without regard to order. The score for recognition is 50 minus twice the number of errors.

*Norms:* (1) Carothers (11) gives the following norms for word recall and word recognition (visual method, twenty-five words, time for study one minute). S's were freshmen college women.

##### Recall

N = 200    Aver. = 11.25    P.E. = 1.74

##### Recognition

N = 200    Aver. = 35.45    P.E. = 5.45

(2) In the Columbia University laboratory, the following results have been obtained for word recall (auditory method, twenty-five words). S's were men and women graduate students.

N = 48    Aver. = 13.65    S.D. = 2.05

## II. LOGICAL MEMORY

### 1. Memory for Ideas or Meanings

In tests of logical memory, the subject is required to reproduce

the ideas or thoughts of a passage of prose which he has read or which has been read to him. Reproduction may be oral, in the form of a narrative; or the subject may be required to give a written account of, or to answer questions upon, the material which he has heard. A well-known test of logical memory occurs at Year 10 in the Stanford-Binet scale. In this passage, which is reproduced in Figure 24, the subject is first instructed to read the passage through. The examiner then covers the passage and asks the subject to tell all that he can remember of what he has read. The "ideas" in the passage are separated by slanting lines. There are twenty-one ideas in all. If a child is able to read the selection in thirty-five seconds with not more than two mistakes, and to recall eight "ideas," he is accredited with an M.A. of 10 years on this test.

Figure 24.—LOGICAL MEMORY TEST AT YEAR 10 (STANFORD-BINET)

New York. / September 5th. / A fire / last night / burned / three houses / near the center / of the city / It took some time / to put it out. / The loss / was fifty thousand dollars, / and seventeen families / lost their homes. / In saving / a girl / who was asleep / in bed, / a fireman / was burned / on the hands.

Several prose selections of the same sort as the one described may be found in Whipple (77). A sample of these, a passage entitled "The Marble Statue," is shown in Figure 25. Age and sex norms for this selection have been published by Pyle (57) and are given in Table XXIV. Pyle's norms apply to written reproduction. The passage is read to the subjects by the experimenter, who then requires each subject to write as much as he can remember. Scoring is in terms of the number of "ideas" correctly reproduced.

Figure 25.—THE MARBLE STATUE—WHIPPLE (77)

A young / man / worked / years / to carve / a white / marble / statue / of a beautiful / girl. / She grew prettier / day by day. / He began to love the statue / so well that / one day / he said to it: / "I would give / everything / in the world / if you would be alive / and be my wife." / Just then / the clock struck / twelve, / and the cold / stone began to grow warm, / the cheeks red, / the hair brown, / the lips to move. / She stepped down, / and he had his wish. / They lived happily / together / for years, / and three / beautiful / children were born / One day / he was very tired, / and grew / so angry, / without cause, / that he struck her. / She wept, / kissed / each child / and her husband, / stepped back / upon the pedestal, / and slowly / grew cold, / pale / and stiff, / closed her eyes, / and when the clock / struck / midnight / she was a statue / of pure / white / marble / as she had been / years before, / and could not hear / the sobs / of her husband / and children.

Logical memory shows a closer relationship to general intelligence than does rote memory. Hurlock and Newmark (31) have reported

TABLE XXIV  
 NORMS FOR "MARBLE STATUE"  
 Number of Ideas Reproduced (Written)  
 (From Pyle [57])

## Boys

Age	8	9	10	11	12	13	14	15	16	17	18	Ad.
No.	102	148	142	149	156	163	129	89	60	45	32	65
Norm	24.3	28.7	30.0	32.9	35.1	36.8	36.1	36.5	34.4	31.6	36.9	38.3
Av. Dev.	6.7	9.1	6.7	5.6	7.1	6.3	7.0	6.7	5.6	8.7	6.0	7.0

## Girls

Age	8	9	10	11	12	13	14	15	16	17	18	Ad.
No.	89	158	138	156	191	161	146	99	91	81	48	86
Norm	28.5	31.0	33.5	36.4	38.1	38.5	39.0	39.1	37.3	36.6	37.8	40.1
Av. Dev.	11.3	9.4	6.8	7.7	7.2	7.1	7.5	6.3	5.1	6.9	4.4	5.9

a correlation of .76 between logical memory for two passages taken from Herring's Revision of the Binet-Simon Scale and I.Q. (Stanford-Binet) in a group of twenty pre-school children. Bolton (5) read a prose selection containing 100 "ideas" to a group of 200 college women. The correlation of this test with the Otis Self-Administering Intelligence Test was .72 (corrected for attenuation). Garrett (19) obtained a correlation of .29 between logical memory for a selection of prose and the Thorndike Intelligence Examination in a group of 158 college students. The students read the prose selection and then answered questions upon it. Although selected to be fairly difficult, the passage turned out to be too easy for the group. Most of the scores were high, and the range was small. Had the selection been more adequate, it is probable that the correlation here would have approximated that found in the other study quoted.

Bolton has reported correlations which range from .20 to .38 between her test of logical memory and five recall and recognition memory tests. Gates (24) obtained correlations of much the same order between logical memory and rote memory. Gates' correlations ranged from .07 to .28 in a group of 197 college students. His rote memory tests included tests of auditory and visual digit span and recognition of geometrical forms. These studies are samples of many others which have found the correlations between logical memory and rote memory to be positive but low.

Logical memory tests have not proved to be highly reliable. Lemmon (39) obtained a reliability coefficient of .60 for a logical memory test consisting of five pages of prose, which the subjects were allowed to study for twenty minutes. Lemmon's group consisted of eighty-eight college men. Bolton obtained a reliability coefficient of .58 for a logical memory test consisting of a passage of easy prose. When this test was rescored for "logical thought" the reliability was increased to .69. This result demonstrates that different methods of scoring, as well as differences in difficulty, have a decided influence upon the reliability of a test.

The tests described in this chapter may be purchased from the C. H. Stoelting Co., Chicago, Illinois.

#### BIBLIOGRAPHY

1. ACHILLES, E. M., "Experimental Studies in Recall and Recognition," *Archives Psychology*, 44, 1920.
2. ANASTASI, A., "A Group Factor in Immediate Memory," *Archives Psychology*, 120, 1930.
3. ANASTASI, A., "Further Studies on the Memory Factor," *Archives Psychology*, 142, 1932.
4. ATKINSON, W. R., "The Relation of Intelligence and of Mechanical Speed to the Various Stages of Learning," *Journal Experimental Psychology*, 12:89-112, 1929.
5. BOLTON, E. B., "The Relation of Memory to Intelligence," *Journal Experimental Psychology*, 14:37-67, 1931.
6. BONSER, F. G., *The Reasoning Ability of Children*, Teachers College, Columbia University, Contributions to Education, 37, 1910.
7. BRONNER, A. F., HEALY, W., LOWE, G. M., AND SHIMBERG, M. E., *A Manual of Individual Mental Tests and Testing*, Little, Brown and Company, Boston, 1927.
8. BURT, CYRIL, "Experimental Tests of General Intelligence," *British Journal Psychology*, 3:94-177, 1909-1910.
9. CALFEE, M., "College Freshman and Four General Intelligence Tests," *Journal Educational Psychology*, 4:223-231, 1913.
10. CAREY, N., "Factors in Mental Processes of School Children," *British Journal Psychology*, 8:70-92, 1915-1917.
11. CAROTHERS, F. E., "Psychological Examinations of College Students," *Archives Psychology*, 46, 1921.
12. CARR, H. A., "The Influence of Visual Guidance in Maze Learning," *Journal Experimental Psychology*, 4:339-417, 1921.
13. CLINTON, R. J., "Nature of Mirror-drawing Ability: Norms on Mirror-

- drawing for White Children by Age and Sex," *Journal Educational Psychology*, 21:221-223, 1930.
14. COLVIN, S. S., "Principles Underlying the Construction and Use of Intelligence Tests," *21st Year Book, National Society for the Study of Education*, Part I, 11-44, 1922.
  15. CONRAD, H. S., AND HARRIS, D., "The Free Association Method and the Measurement of Adult Intelligence," *University of California Publications in Psychology*, 5:1-45, 1931.
  16. DEARBORN, W. F., AND LINCOLN, E. A., "A Class Experiment in Learning," *Journal Educational Psychology*, 13:330-340, 1922.
  17. FISHER, V. E., "A Few Notes on Age and Sex Differences in Mechanical Learning," *Journal Educational Psychology*, 18:562-564, 1927.
  18. FOSTER, W. S., AND TINKER, M. A., *Experiments in Psychology*, Henry Holt and Company, Inc., New York, 1929. (Notes for Instructors.)
  19. GARRETT, H. E., "The Relation of Tests of Memory and Learning to Each Other and to General Intelligence in a Highly Selected Group," *Journal Educational Psychology*, 19:601-613, 1928.
  20. GARRETT, H. E., *Statistics in Psychology and Education*, Longmans, Green and Co., New York, 1926.
  21. GARRISON, K. C., *An Analytic Study of Rational Learning*, George Peabody College for Teachers, Contributions to Education, 44, 1928.
  22. GARRISON, K. C., "Further Studies in Various Types of Speed Performances as Related to Mental Ability," *Journal Genetic Psychology*, 36:344-349, 1929.
  23. GATES, A. I., "Correlations of Immediate and Delayed Recall," *Journal Educational Psychology*, 9:439-496, 1918.
  24. GATES, A. I., "Correlations and Sex Differences in Memory and Substitution," *University of California Publications in Psychology*, 1:345-350, 1916.
  25. GATES, A. I., "Variations in Efficiency During the Day, Together with Practice Effects, Sex Differences and Correlations," *University of California Publications in Psychology*, 2:1-156, 1916.
  26. GOODENOUGH, F., *The Kuhlmann-Binet Tests*, University of Minnesota Press, Minneapolis, 1928.
  27. HALLOWELL, D. K., "Mental Tests for Pre-school Children," *Psychological Clinic*, 16:235-276, 1925-1928.
  28. HARTMANN, G. W., "The Relative Influence of Visual and Auditory Factors in Spelling Ability," *Journal Educational Psychology*, 22:691-699, 1931.
  29. HAUGHT, B. F., "The Interrelation of Some Higher Learning Processes," *Psychological Monographs*, 30:6, 1921.
  30. HOLLINGWORTH, H. L., "Articulation and Association," *Journal Educational Psychology*, 6:99-105, 1915.
  31. HURLOCK, E. B., AND NEWMARK, E. D., "The Memory Span of Pre-school Children," *Journal Genetic Psychology*, 39:157-173, 1931.

32. HUSBAND, R. W., "A Comparison of Human Adults and White Rats in Maze Learning," *Journal Comparative Psychology*, 9:361-377, 1929.
33. JASTROW, J., "Community of Ideas of Men and Women," *Psychological Review*, 3:68-71, 426-431, 1896.
34. JUNG, C. J., "The Association Method," *American Journal Psychology*, 21:219-269, 1910.
35. KENT, G. H., AND ROSANOFF, A. J., "A Study of Association in Insanity," *American Journal Insanity*, 67:37-96; 317-390, 1910-1911.
36. KNOTTS, J. R., AND MILES, W. R., "Notes on the History and Construction of the Stylus Maze," *Journal Genetic Psychology*, 35:415-427, 1928.
37. LANDIS, C., GULLETTE, R., AND JACOBSON, C., "Criteria of Emotionality," *Pedagogical Seminary and Journal Genetic Psychology*, 32:209-234, 1925.
38. LEACH, H. M., AND WASHBURN, M. F., "Some Tests by the Association Reaction Method of Mental Diagnosis," *American Journal Psychology*, 21:162-167, 1910.
39. LEMMON, V. W., "The Relation of Reaction Time to Measures of Intelligence, Memory, and Learning," *Archives Psychology*, 94, 1927.
40. MANCHESTER, G. S., "Experiments on the Unreflective Ideas of Men and Women," *Psychological Review*, 12:50-66, 1905.
41. MCGEOCH, G. O., "The I.Q. as a Factor in the Whole-part Problem," *Journal Experimental Psychology*, 14:333-358, 1931.
42. MILES, W. R., "The High-relief Finger Maze for Human Learning," *Journal General Psychology*, 1:3-14, 1928.
43. MUNSTERBERG, H., *On the Witness Stand*, The McClure Company, New York, 1908.
44. MURPHY, G., "Types of Word-association in Dementia Præcox, Manic-depressive, and Normal Persons," *American Journal Psychiatry*, 2:539-571, 1923.
45. NÖH, E. J., AND GUILFORD, J. P., "Sex Differences and the Method of Continuous Lists," *American Journal Psychology*, 42:415-419, 1930.
46. NORSWORTHY, N., "The Psychology of Mentally Deficient Children," *Archives Psychology*, 1, 1906.
47. NYSWANDER, D. B., "A Comparison of the High-relief Finger Maze and the Stylus Maze," *Journal General Psychology*, 2:273-289, 1929.
48. OTIS, M., "A Study of Association in Defectives," *Journal Educational Psychology*, 6:271-288, 1915.
49. PATERSON, D. G., ELLIOTT, R. M., *et al.*, *Minnesota Mechanical Ability Tests*, University of Minnesota Press, Minneapolis, 1930.
50. PETERSON, J., "The Backward Elimination of Errors in Mental-maze Learning," *Journal Experimental Psychology*, 3:257-280, 1920.
51. PETERSON, J., "Comparison of White and Negro Children in the Rational Learning Test," *27th Yearbook, National Society for the Study of Education*, 333-341, 1928.

52. PETERSON, J., "Experiments in Rational Learning," *Psychological Review*, 25:443-467, 1918.
53. PETERSON, J., "The Rational Learning Test Applied to Eighty-one College Students," *Journal Educational Psychology*, 11:137-150, 1920.
54. PETERSON, J., AND ALLISON, L. W., "Effects of Visual Exposure on the Rate and Reliability of Stylus-maze Learning," *Journal General Psychology*, 4:36-48, 1930.
55. PETERSON, J., AND LANIER, L. H., "Studies in the Comparative Abilities of Whites and Negroes," *Mental Measurement Monographs*, 5, 1929.
56. PORTEUS, S. D., *Porteus Tests, The Vineland Revision*, Publications of the Training School at Vineland, New Jersey, 16, 1919.
57. PYLE, W. H., *The Examination of School Children*, The Macmillan Company, New York, 1913.
58. PYLE, W. H., *Laboratory Manual in the Psychology of Learning*, Warwick and York, Baltimore, 1923.
59. PYLE, W. H., *The Nature and Development of Learning Capacity*, Warwick and York, Baltimore, 1925.
60. ROSANOFF, A., *Free Association Test (Kent-Rosanoff)*, Reprinted from *Manual of Psychiatry*, John Wiley and Sons, sixth edition, 1927.
61. ROSANOFF, I. R., AND ROSANOFF, A. J., "A Study of Association in Children," *Psychological Review*, 20:43-89, 1913.
62. RUCH, G. M., "Influence of the Factor of Intelligence on the Form of The Learning Curve," *Psychological Monographs*, 34:7, 1925.
63. SCHNECK, M. M. R., "The Measurement of Verbal and Numerical Abilities," *Archives Psychology*, 107, 1929.
64. SMITH, W. W., *The Measurement of Emotion*, Harcourt, Brace and Company, New York, 1922.
65. SNODDY, G. S., "An Experimental Analysis of Trial-and-Error Learning in the Human Subject," *Psychological Monographs*, 124, 1920.
66. SPEARMAN, C., *Abilities of Man*, The Macmillan Company, New York, 1927.
67. SPENCE, K. W., AND TOWNSEND, S., "A Comparative Study of Groups of High and Low Intelligence in Learning a Maze," *Journal General Psychology*, 3:113-130, 1930.
68. STARCH, D., "A Demonstration of the Trial-and-Error Method of Learning," *Psychological Bulletin*, 7:20-23, 1910.
69. STARR, A. S., "The Diagnostic Value of the Audito-vocal Digit Memory Span," *Psychological Clinic*, 15:61-84, 1923-1924.
70. TERMAN, L. M., *Measurement of Intelligence*, Houghton Mifflin Company, New York, 1916.
71. TERMAN, L. M., *Stanford Revision and Extension of the Binet-Simon Scale*, Warwick and York, Baltimore, 1917.
72. TERMAN, L. M., "The Vocabulary Test as a Measure of Intelligence," *Journal Educational Psychology*, 9:452-466, 1918.
73. TERMAN, L. M., AND CHILDS, H. G., "A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence," *Journal Educational Psychology*, 3:198-208, 1912.

74. WARDEN, C. J., "The Relative Economy of Various Modes of Attack in the Mastery of a Stylus Maze," *Journal Experimental Psychology*, 7:243-275, 1924.
75. WEIDENSALL, J., *The Mentality of the Criminal Woman*, Warwick and York, Baltimore, 1916.
76. WHEAT, L. B., *Free Association to Common Words*, Teachers College, Columbia University, Contributions to Education, 498, 1931.
77. WHIPPLE, G. M., *Manual of Mental and Physical Tests, Part II*, Warwick and York, Baltimore, 1915.
78. WHITLEY, M. T., "An Empirical Study of Certain Tests for Individual Differences," *Archives Psychology*, 19, 1911.
79. WILLOUGHBY, R. R., "Incidental Learning," *Journal Educational Psychology*, 20:670-682, 1929.
80. WISSLER, C., "The Correlation of Mental and Physical Traits," *Psychological Monographs*, 16, 1901.
81. WOODROW, H., AND LOWELL, F., "Children's Association Frequency Tables," *Psychological Monographs*, 22:97, 1916.
82. WOODWORTH, R. S., AND WELLS, F. L., "Association Tests," *Psychological Monographs*, 13:57, 1911.
83. WYATT, S., "The Quantitative Investigation of Higher Mental Processes," *British Journal Psychology*, 6:109-133, 1913-1914.
84. WYLIE, A. T., *The Opposites Test*, Teachers College, Columbia University, Contributions to Education, 170, 1925.
85. YOAKUM, C. S., AND CALFEE, M., "An Analysis of the Mirror-drawing Experiment," *Journal Educational Psychology*, 4:283-292, 1913.

## THE PSYCHOGRAPH

The psychograph is a relatively simple and direct method of representing an individual's scores in a number of tests. One type of psychograph is shown in Figure 26. This diagram represents the scores made by a woman graduate student, B.S., upon eleven tests like those described in Part I of this book. The method of constructing this figure was to convert all of the scores into "sigma" or "standard" scores. This was done in the following manner. Suppose an individual obtains a score of 62 on a test in which the mean score is 48.25 and the S.D. 8.53; and a score of 122 in a test in which the mean score is 138.42 and the S.D. 25.46. To convert these scores into sigma values we subtract the mean score on each test from the subject's score in that test and divide by the standard deviation, *e.g.*:

$$\frac{62 - 48.25}{8.53} = +1.61$$

$$\frac{122 - 138.42}{25.46} = - .64$$



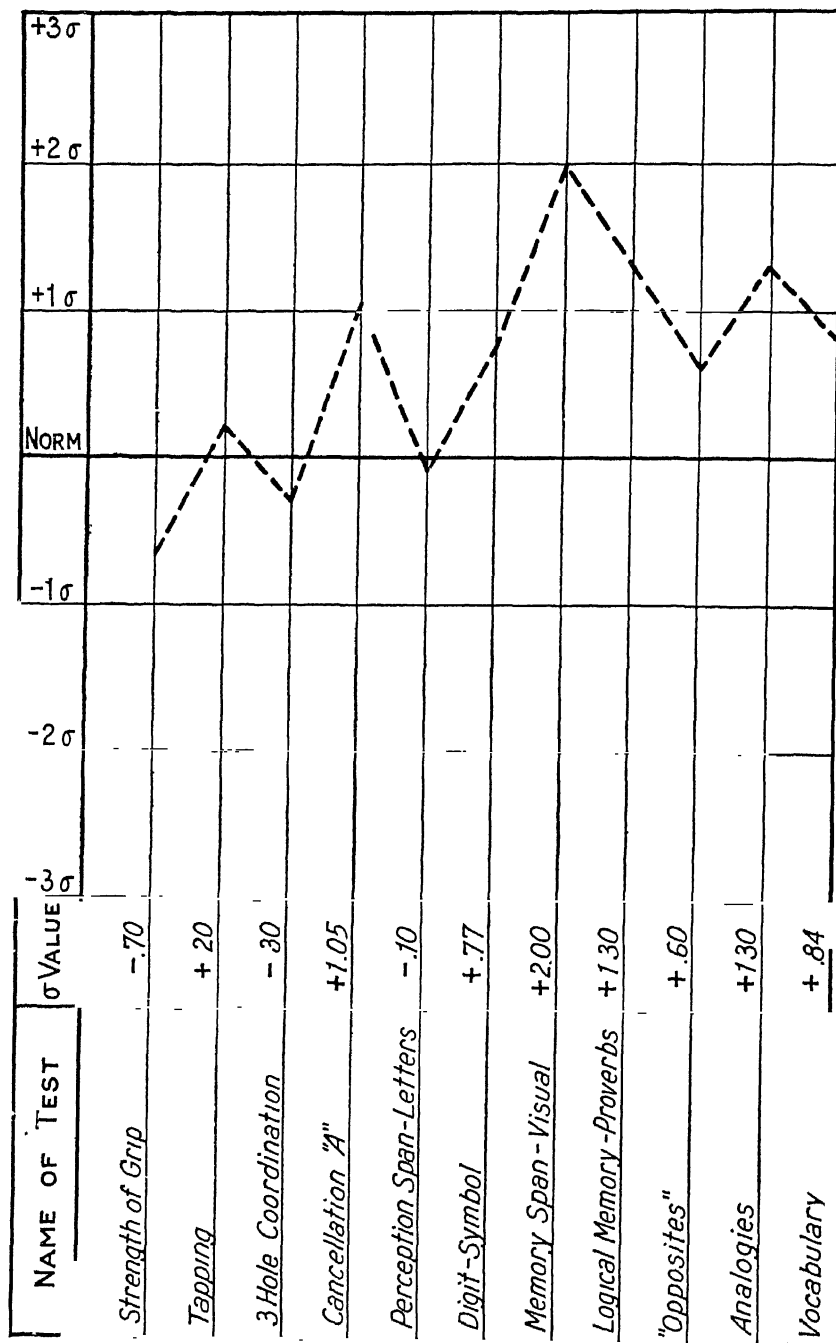


Figure 26.—PSYCHOGRAPH

A sigma score is *plus* when the subject's score is above the mean; and *minus* when the subject's score is below the mean. The student must remember, however, that when scores are in terms of time (seconds) a score *above* the mean will give a minus sigma score, and one *below* the mean a plus sigma score.

There are several distinct advantages in the use of a psychograph. In the first place, it enables one to compare scores which are expressed in very different units (20). It would be impossible to compare directly the number of taps in thirty seconds with the size of one's vocabulary. Secondly, a psychograph enables one to get a compact picture of an individual's relative achievements in different functions. In Figure 26, for instance, it will be at once noted that B.S. ranks about average upon those tests which deal with simple sensory-motor functions, but is above the mean or norm in learning, memory and language ("general intelligence") tests. From the psychograph alone, the most obvious conclusion would be that B.S. is "bright" (in an academic sense) and somewhat less advanced in motor and sensory ability.

Two sheets are provided on pp. 138, 139, upon which the student may construct a psychograph of his own abilities based upon a selection of the tests described in Part I.

### ERRATUM

The graph paper for making the psychograph will be found following page 224 of Part Two.



■

P A R T T W O

■



## CHAPTER I

### VERBAL OR LINGUISTIC TESTS OF "GENERAL INTELLIGENCE"

THE task of this chapter is to describe certain groups or batteries of tests which present a variety of problems—cutting across many mental processes—in the attempt to measure general ability or “general intelligence.” When these tests demand, as essential to their performance, a knowledge and understanding of language and numbers, they may be more precisely described, perhaps, as verbal or linguistic tests of “general intelligence.” It will be the position taken in this chapter that at least for children such tests are, in the main, measures of “general” scholastic aptitude, *i.e.*, of the ability to perform well the work of the school. For adults, they are measures of the ability to perform such occupational tasks as depend mainly upon schooling. The evidence for this view will accumulate as we go through the chapter. Suffice it to say that it comes chiefly from an analysis of the “general intelligence” test itself, and from experimental studies in which such tests have been employed.

The term “general intelligence” has been taken by most psychologists to cover a much wider range of behavior activities than those ordinarily exercised in educational acquisition (58). Thus, Binet (1) defined intelligence as (1) the tendency to take and maintain a definite direction; (2) adaptability to new situations and new requirements; and (3) the ability to criticize one’s own acts. William Stern (56) considered an intelligent person to be one who could adjust his thinking to new requirements, *i.e.*, was mentally adaptable. Terman (58) has defined intelligence as being essentially the ability to carry on abstract thinking, while Colvin (58) has emphasized ability to learn easily and quickly. Woodworth (75) has offered the following analysis: Intelligent activity, he says, is characterized (1) by retentivity, or the ability to use facts and activities already acquired; (2) by ready adaptability to novel situations; (3) by curi-

osity, interest in, and desire to know about things; and (4) by persistence, the trait of sticking to what one begins.

All of these definitions emphasize adaptability to life situations, except Terman's, which is close to the view taken in this chapter. Any definition of *general* intelligence must of necessity be broad and somewhat loose, because of the very nature of the function itself. And to measure such a comprehensive function becomes a difficult, if not an impossible, task. In order to give greater precision to the concept of general intelligence, Thorndike (66) has suggested that at least three levels or stages of intelligent activity be recognized. These "intelligences" he calls the abstract, the mechanical and the social. Abstract intelligence is "the ability to understand and manage ideas and symbols, such as words, numbers, chemical or physical formulæ, legal decisions, scientific principles, and the like. . . ." This is practically the equivalent of what we have called scholastic aptitude. Mechanical intelligence includes "the ability to learn, understand, and manage things and mechanisms, such as a knife, a gun, a mowing machine, automobile, boat, lathe." Social intelligence is "the ability to understand and manage men and women, boys and girls, to act wisely in human relations." According to Thorndike, we should expect social intelligence to be found at a high level in salesmen, politicians, and clerks; mechanical intelligence in carpenters, masons, and plumbers; and abstract intelligence in teachers, scholars, and scientists. Presumably, the "generally intelligent man" would rank relatively high on all three levels.

While the makers of general intelligence tests have, for the most part, constructed their tests to measure ability on what Thorndike has called the abstract level, most of them have not hesitated to call their products measures of general intelligence. There are probably several reasons for this. One grows out of the common tendency to think of intelligence as a faculty or entity (31). Another is the fact that most test makers have been educators to whom, naturally enough, intelligence is closely synonymous with academic achievement. General intelligence tests have been most widely, and successfully, used in colleges and schools (p. 50). When taken over into business and industry, such tests have been by no means so successful, especially when employed to predict aptitude in occupations requiring manual skill and rather definite personality traits (7). The correlations of general intelligence tests with mechanical aptitudes (Chapter II) as

well as with the various social traits (Chapter III) have proved to be low or negligible. Hence, the term general intelligence test as applied to verbal or linguistic test batteries would seem to be clearly a misnomer, if such tests are to be taken as measures of the general ability, defined on p. 3. The term has become so well established in mental-testing vocabulary, however, that we shall continue to use the term "general intelligence tests." We shall, however, frequently remind the student of the more precise meaning of the term by using the qualifying adjective "verbal" as applied to such tests.

### TESTS OF GENERAL INTELLIGENCE

Tests of general intelligence on the verbal or abstract level fall into two main groups: *oral* or *individual tests* and *written* or *group tests*. Individual tests are administered privately to one child at a time and require from thirty minutes to an hour per person. These scales, while largely linguistic at the higher age levels, involve many activities of a manipulative and non-language sort at the lower ages. For young children, therefore, they are not measures of abstract or verbal ability to such a degree as for older pupils. (See Chapter II.)

Group or written tests of intelligence are much like the ordinary school examination. Usually they consist of a booklet containing from three to ten separate sub-tests, which are administered with definite time limits for each. Group tests demand facility with language and ability in calculation; a fund of common information; memory for what one has learned; and ingenuity in seeing relations and in using symbols in the solution of problems. These are much the same abilities called out in school work. In fact, Kelley (33) has shown that ordinarily about 90 per cent. of the general intelligence test and the all around school achievement test measures the same functions.

### ORAL OR INDIVIDUAL TESTS OF INTELLIGENCE

The individual test of general intelligence, as we know it today, grew out of the work of Alfred Binet (2). Faced with the practical problem of identifying feeble-minded and low-grade children in the Paris schools, Binet, in 1905, in collaboration with the physician, Thomas Simon, devised a series of mental tests to be used with children. These tests were revised and extended in 1908 and again in 1911.



Binet's tests marked a real development in the measurement of general ability for two reasons: first, they were devised to gauge complex mental processes such as foresight, reasoning, and judgment, instead of simple, sensory-motor functions, such as speed and accuracy of movement; and, secondly, Binet first introduced the idea of an age-scale. In an age-scale all of the tests intended for a given age are grouped together. Tests for seven-year-olds are placed in the seven-year group, those for eight-year-olds in the eight-year group, and so on. If a child of eight is able to do correctly all of the tests up to and including his own age group, he earns a mental age of eight years. This means that so far as the age-scale is able to measure it, he has attained a mental development characteristic of the average child of eight years. Should a child of eight pass the tests up to and including those of the ten-year level, his mental age is ten years, or two years in advance of his life or chronological age. On the other hand, if a child is retarded intellectually one or more years, his mental age, as found from the scale, will fall below his chronological age.

Binet's scale, as he left it in 1911, contained fifty-four tests in all. These were arranged to span the age levels from three to fifteen years. Binet's plan was to regard a test which could be passed by from 60 to 90 per cent. of a given age group as suitable for that age. Five tests were placed in each age group from three to ten (except at age four where there were only four tests); five tests were placed at age twelve, and five at age fifteen. Five tests were also set up to constitute an adult level.

While Binet's scale was in a real sense a departure from previous work in mental measurement, it was at the same time definitely shaped by two trends then current in psychology. In the first place, there was the need of securing reliable measures of degrees of feeble-mindedness; and in the second place, there was an increasing interest in the problem of individual differences. The history of both of these movements has been given in detail elsewhere and need not be repeated here (45). Suffice it to say that Binet's work was focused on the one hand by the clinical studies upon the feeble-minded of Pinel, Itard, and Seguin in France and Kraepelin in Germany; and on the other hand, by the statistical studies of Galton and Pearson in England upon family resemblances and the influence of heredity and environment upon achievement.

## REVISION OF BINET'S TESTS IN AMERICA

Several revisions of Binet's tests have appeared in America: that of Goddard (1911), that of Kuhlmann (1912 and 1922), the Point Scale by Yerkes, Bridges, and Hardwick (1915 and 1923), the Stanford Revision of Terman (1916), and the revision by Herring (1922). Goddard's revision and the point scale will be considered briefly. The first is chiefly of historic interest, while the point-scale method has been largely supplanted in individual testing by the age-scale. The Terman, Kuhlmann, and Herring Revisions are today the most widely used individual scales for the measurement of general intelligence. For this reason these tests will be described in some detail.

**1. Goddard's Revision of the Binet-Simon Tests (1911)**

Goddard's (26) revision of Binet's tests was the first translation and adaptation of this scale to be extensively used in America. Goddard shifted the age location of several of the tests; introduced a few new tests into the fifteen-year group; and adapted the terminology and content of the scale for use with American children. This scale was widely used until the appearance of the Stanford Revision in 1916.

**2. The Yerkes-Bridges-Hardwick Point Scale (1915)**

In the Yerkes-Bridges-Hardwick Point Scale (77) the separate tests are arranged in an ascending order of difficulty, instead of being grouped at age levels as in an age-scale. Credit points are assigned for passing each test in the scale, the number of points depending upon the number of items in the test. Thus, the total score for the memory span for digits test, which consists of five items, is five points. The Yerkes-Bridges-Hardwick Point Scale, while it makes use of Binet's tests, differs from Binet's scale both in method and scoring. These authors objected to Binet's plan of grouping tests into age groups, and also to his method of scoring a test as either right or wrong, without allowing partial credits. The Point Scale consists of twenty tests, of which nineteen were taken from Binet. A child taking the test is given so many points for each test passed, partial credits being allowed according to the number of items in each test. The total number of points earned out of a possible maximum of 100 is the basis for the child's mental rating. If a boy of nine whose

native language is English scores fifty-six points, for example, his mental age is nine years, as fifty-six is the norm for English-speaking nine-year-old boys. As a measure of relative standing, or brightness, the authors employed a Coefficient of Intellectual Ability (C.I.A.), which is obtained by dividing a child's earned score by the standard score or norm for his chronological age. In the illustration given above, the C.I.A. of our nine-year-old boy would be  $56/56$ , or 1.00. If, however, this child had been twelve years old, the standard score or norm for which age is seventy-five points, his C.I.A. would have been  $56/75$ , or .75. Separate norms or standard scores have been calculated for English- and non-English-speaking children, for the sexes, and for good and poor social environment, since the authors contend that a child's rating is influenced by all of these factors.

It is possible to change scores from the Point Scale into mental-age equivalents which have much the same meaning as the M.A. obtained from an age-scale. A correlation of .87 in a group of about 300 female delinquents (adults) has been obtained between scores on the Point Scale and Stanford-Binet I.Q.'s (19).

In 1923, a revision of the Point Scale was undertaken by Yerkes and Foster (76). Only a few slight changes were made in the original scale. However, an infant scale, consisting of twenty-two tests, was added for children seven years old and below, and an adolescent-adult scale of twenty items for those above thirteen. These extensions were made because the Point Scale, although intended for children from three to sixteen years old, had proved to be most effective in the age range from seven to thirteen years. The new scales are still somewhat provisional, however, their standardization being as yet tentative.

### 3. Stanford Revision of the Binet-Simon Tests (1916)

The Stanford Revision of the Binet-Simon Tests is the most widely used age-scale (60, 64). In many respects, too, it is the most valuable individual intelligence test. This revision was made by L. M. Terman and his associates at Stanford University, from which the scale gets its name. The reasons for undertaking a revision of Binet's tests are given by Terman as follows: (1) the original tests were too few and too difficult at the upper age levels and too easy at the lower age levels; (2) the directions for giving the tests were often in-

definite; (3) many tests were misplaced in the scale, some being too far up and others too far down.

Two scores are obtained from the Stanford Revision. The first is the Mental Age, or M.A., and the second is the Intelligence Quotient, or I.Q. The I.Q. is the mental age divided by the chronological age, and is a measure of brightness, as contrasted with the mental age, which is a measure of mental status. The two measures are supplementary, each giving distinctive information. Thus, a child of eight years and a man of forty may each have a mental age of eight years, *i.e.*, be of the same intellectual status as far as the tests are concerned. But the child has an I.Q. of 100 (8/8) and is normal, while the man is feeble-minded, with an I.Q. of 50 (8/16) (p. 24).

#### 4. Characteristics of the Stanford Revision

The most important facts relating to the Stanford Revision, or Stanford-Binet, as it is often called, may be summarized as follows:

(a) *Data*: There were 2,300 subjects in all, consisting of 1,700 normal children, 200 defective and superior children and 400 adults.

(b) *Method*: (1) All of the data on each test for each age group were assembled, *i.e.*, the per cent. passing, comments and notes of examiners, *etc.*, and three provisional scales drawn up before the final revision was made; (2) forty new tests in addition to the original Binet tests were tried out, the plan being to have six tests, valued at two months each, placed at each age level; (3) the positions of the tests in the scale were made to depend upon the answers of approximately 1,000 native-born children of average social status, five to fourteen years old, and each within two months of a birthday; (4) to secure uniform results one-half year was spent in training examiners and another half year in supervising the giving of tests.

(c) *Placement of Tests*: The ultimate goal was to secure an arrangement of tests at each age which would make the median mental age coincide exactly with the median chronological age. Thus, the average child of six should test exactly at age six on the scale; the average child of eight exactly at age eight, *etc.*; or, to say the same thing differently, the median intelligence quotient at each age should be 100. When the distributions, in the preliminary scoring, gave a median I.Q. above or below 100 for a given age group, the tests were shifted up or down, or the scoring changed, until the median M.A. equaled the median C.A.

(d) *Validity of Tests*: To be included in the scale, each test had to show an increase in the percentage of children passing it from one year to the next. All of the children tested were divided into three groups on the basis of total score: those below 90 I.Q.; those 90 to 109 inclusive;

and those 110 and above. Each test was then examined to see if it showed a decidedly higher per cent. of passes from the inferior to the superior group, and only those tests which passed this criterion were considered adequate. A high correlation between I.Q.'s and quality of school work was set up as a third criterion. School achievement and teachers' ratings were considered to be valuable indices of intellectual ability, to which any intelligence scale must be related.

(e) *Reliability of Tests.* The reliability of the Stanford-Binet scale as measured by its self-correlation will vary, of course, with the size and character of the group, but it ranges usually from .80 to .95 (33). The P.E. of an I.Q. is approximately three and one-half points, which means that at least 50 per cent. of the children tested by the scale will be within three and one-half points of their true I.Q.'s (44).

(f) *The Final Revision.* In its final form, the Stanford-Binet contained ninety tests, thirty-six more than the Binet scale. Six tests, each valued at two months, were placed at the ages from three to ten; eight tests valued at three months each were placed at age twelve; six tests valued at four months each at age fourteen; six tests valued at five months each at age sixteen, the average adult level, and six tests valued at six months each at age eighteen, the superior adult level. Alternate tests to be used when, for some reason, one of the regular tests did not apply were placed at each age level, except at age twelve and superior adult.

## 5. The Kuhlmann Revision of the Binet Tests (1922)

Kuhlmann's first revision of the Binet-Simon tests was published in 1912 (37), and followed closely the original scale. His second revision, in 1922 (36), which is the one usually considered, made the following changes and additions:

(a) Nineteen tests judged to be unsatisfactory were eliminated. These tests, such as naming the days of the week, counting thirteen pennies, giving the age, writing and reading were considered to depend too much upon variable factors of training to be good intelligence tests.

(b) Scoring was made more objective, in terms of errors and time wherever possible, in order to exclude the personal element of judgment.

(c) The scale was extended down to the three months' level. Five tests were placed at three months; five at six months; five at Year 1; five at eighteen months; and five at Year 2. The maximum mental age was set at fifteen years instead of at sixteen as in the Stanford-Binet.

(d) Above the two-year level the number of tests at each age

was increased to eight. This was to allow more flexibility in the use of abbreviated forms of the scale.

(e) The revision was based on the records obtained from some 7,000 children tested over a period of seven years.

As in the Stanford-Binet, two scores are obtained from the Kuhlmann Revision, the M.A. and the I.Q. With normal children, these measures have practically equivalent meanings in the two scales. The Kuhlmann Revision has been widely used with older children, but has found its chief use with very young children. The reliability of this scale as determined from retests upon 300 children, two to four years of age, was  $.81 \pm .01$  (27).

#### 6. The Herring Revision of the Binet Tests (1922)

The Herring Revision (29) is a point scale made up of tests adapted for the most part from the Binet and the Stanford-Binet Scales. Several new tests were also added. In all there are thirty-eight tests in this scale, most of them involving language and fairly complex verbal relations. The scale is subdivided into five groups: tests 1-4 called Group A; tests 1-13, Group B; tests 1-22, Group C; tests 1-31, Group D; and tests 1-38, Group E. A score can be obtained from each group of tests separately. Tables have been provided from which the points scored may be translated into mental-age equivalents. A mental age can be determined from Group A alone, or from Group B, C, D, or E. The score in the first group determines the tests which are to be given in successive groups, thus making it unnecessary to administer all of the tests to a given subject.

The Herring Revision will prove valuable as an alternate scale to be used in place of the Stanford Revision. The scale was validated against Stanford-Binet and ratings from it have much the same value as Stanford-Binet M.A.'s and I.Q.'s. The correlation between Stanford-Binet M.A.'s and Herring Group E M.A.'s (*i.e.*, whole scale) is reported by the author to be .98, in a group of 116 children, all twelve years of age; and .99 in a group of 154 children, four to eighteen years old. In the case of very bright children, Herring-Binet M.A.'s and I.Q.'s do not correspond so closely to Stanford-Binet values as they do for average children (8). The reliability of the Herring-Binet Scale (Group E I.Q.'s) was .99 in a group of eighty-two children.

For the most accurate results it is advisable to give the Herring-

Binet through Groups D or E. This scale has the advantage of being one of the easiest individual tests to administer. Instructions for giving and scoring the separate tests as well as directions for obtaining M.A.'s from point-score values are clearly set forth in the manual, *Herring Revision of the Binet-Simon Tests, Examination Manual Form A*, World Book Company, Yonkers, New York.

#### ADMINISTRATION OF THE STANFORD AND KUHLMANN REVISIONS

##### 1. The Stanford Revision of the Binet Tests

Full directions for giving the Stanford Revision of the Binet-Simon Scale as well as a discussion of the meaning and value of the separate tests will be found in Terman's book, *The Measurement of Intelligence* (60). For those who are well acquainted with the test and need only the directions in a simple form, *The Condensed Guide for the Stanford Revision of the Binet Tests* will be useful. Test materials and booklets for recording answers may be purchased from the Houghton Mifflin Company, New York City.

In scoring the Stanford-Binet Scale the Examiner first locates the "basal age," that is, the age at and below which all of the tests are passed. Additional credits obtained by passing the tests at successive year levels are then combined with the basal age to give the final M.A. rating. Ordinarily, there are six tests at a given age level, each test having a value of two months on the age scale. If there are fewer than six tests, as there are at the upper year levels, each test is given proportionally greater weight. To illustrate the scoring procedure, if a child passes all of the tests at Year 8 and below, three at Year 9, and one at Year 10, his M.A. is 8 years (basal age) +  $3 \times 2$  months, +  $1 \times 2$  months, or 8 years and 8 months. If the child's chronological age (C.A.) is 8 years 2 months, his I.Q. is 106, i.e.,

$$\frac{104 \text{ mos.}}{98 \text{ mos.}}$$

The maximum mental age which can be achieved on the Stanford-Binet is 19.5. This M.A. is obtained when an individual passes all of the sixteen-year tests, earning a basal age of 16.5<sup>1</sup> years, and then passes all of the eighteen or superior adult tests. There are six tests at the eighteen-year or superior adult level, each carrying a

<sup>1</sup>There are six tests at Year 16, each with a value of five months (total 25 years). If an individual passes all the tests at Year 14, and then, in addition, all of the tests at Year 16, his M.A. is  $14 + 2.5$  or 16.5 years.

credit of six months' M.A. Hence passing all six of these tests in addition to those at Year 16 gives the individual 3.0 additional years' credit, or a total of  $16.5 + 3.0$  or 19.5 years.

No one should attempt to administer the Stanford Revision until he has had at least one course in mental testing preceded by a substantial background in psychology. This requirement is doubly necessary because the tests seem so easy to give. The purpose of a mental test is to discover, not what a child can do when prodded or aided by the examiner, but what he can do under strictly controlled conditions. The trained examiner knows the directions for giving the tests "by heart"; he studies the child's reactions and not his instruction book. Furthermore, and more important, he knows how to interpret his results. Children's abilities are not evenly developed. One child will do particularly well on a memory test, another on tests involving numbers, still another on tests of language and vocabulary. Again, two children may earn the same mental age, but the one will "scatter" widely up and down the scale, while the other passes few tests above his final level. All of these idiosyncrasies are important and significant. There is no reason why the average teacher cannot learn to give the Stanford Revision accurately. But to do so she must lay aside her rôle as teacher and become an examiner. Also, she must be technically prepared for the work. The well-intentioned but untrained amateur does far more harm than good when she administers and tries to interpret a test which she does not fully understand.

## 2. The Kuhlmann Revision

Directions for giving the Kuhlmann Revision, instructions as to the conduct of the examination, rules for finding an I.Q., and tables to facilitate the calculation of I.Q.'s are contained in Kuhlmann's *A Handbook of Mental Tests* (36). Test materials may be purchased from Warwick and York, Inc., Baltimore, Maryland.

Even more than in the case of the Stanford-Binet is it necessary for one to have training before attempting to administer the Kuhlmann Revision. The material employed is considerably more detailed and complex than in the Stanford-Binet, and the calculation of the I.Q. is more involved. Since this test is used widely with young children, the examiner must know directions and procedure thoroughly. It seems probable that the author intentionally made his testing technique intricate in order to discourage untrained examiners.



## CHARACTERISTICS OF THE AGE-SCALE

## 1. Meaning and Significance of the M.A. and I.Q.

Since it is a measure of mental *status*, mental age will ordinarily change with advances in chronological age. As a child grows older, his mental age keeps pace with his chronological age if he is normal or average in mental ability, exceeds it if he is advanced, and falls behind it if he is retarded. The effect of an age increase upon the I.Q. is very different from its effect upon mental age. The I.Q. is a measure of relative *brightness*, and hence when a child's position in his age group has once been defined by his I.Q., this relationship must be maintained as age increases if the I.Q. is to be of value. If, for example, a boy has been classified as bright, normal, or dull, he must retain this classification at succeeding ages, since a scale would obviously be quite useless if the same child tested bright at age eight, normal at age nine, and dull at age ten. The question of whether the I.Q. remains constant with increasing age becomes exceedingly important when we are dealing with age-scales like the Kuhlmann or Stanford-Binet. There are two angles from which the problem may be attacked: the one theoretical and the other practical. First, does mental progress, as measured by the age-scale, follow a growth curve which will give a constant I.Q.? Secondly, does the I.Q. calculated from the age-scale really turn out to be constant in actual practice? We shall consider these two problems in order.

## 2. Theoretical Requirements for a Constant I.Q.

Suppose that a child of five earns an M.A. of six on the Stanford-Binet Scale, giving him an I.Q. of 120 ( $6/5$ ). Five years later, when this child is ten years old, his M.A. must be twelve if his I.Q. is to remain at 120. In other words, in order for this child's I.Q. to remain constant he must be *two* years advanced at age ten if he was *one* year advanced at five. This means that acceleration in mental growth is twice as great at year ten as at year five, or else that the mental age unit *decreases* as the child grows older, so that two years of mental growth at age ten equals one year of mental growth at age five. In order to decide which of these alternatives fits the Binet scale, we shall investigate the characteristics of the mental growth curve over the interval spanned by the Binet tests.

Figure 1 shows a hypothetical mental growth curve which will give a constant I.Q. Two characteristics of this age-progress curve

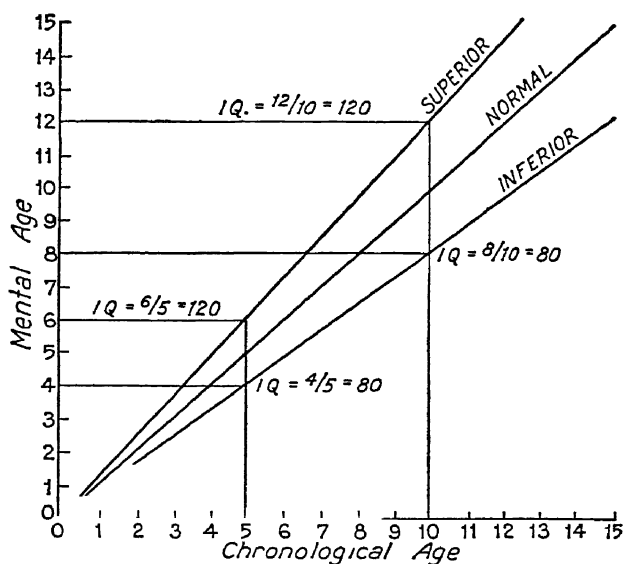


Figure 1.—HYPOTHETICAL GROWTH CURVES WHICH GIVE A CONSTANT I.Q.

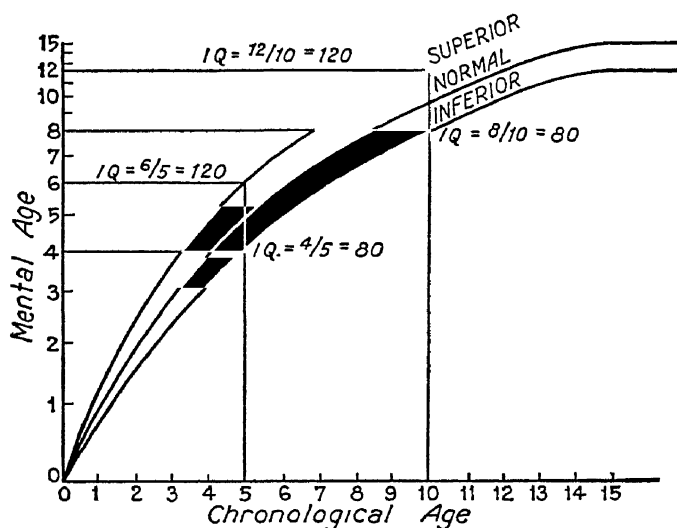


Figure 2.—HYPOTHETICAL GROWTH CURVES WHICH GIVE A CONSTANT I.Q.

must be carefully noted. First, the M.A. is represented as a constant unit on the vertical axis, the C.A. being shown as a constant unit on the base line. Secondly, the "curve" of normal or average growth is a straight line from which the "curves" of superior and inferior ability (also straight lines) diverge more and more as time goes on. This regular and increasing divergence of the upper and lower growth lines—represented by an increasing spread (SD) of the test scores at succeeding ages—accounts for the constant I.Q. To illustrate, consider the hypothetical child whose mental age at age five is six, his I.Q. being  $6/5$ , or 120. As shown by the upper line of growth, the I.Q. of this child will also be 120 at age ten, provided his rate of growth is maintained, his M.A. at age ten being twelve years. Similarly, a child whose I.Q. at age five is  $80 (4/5)$  will have an I.Q. of  $80 (8/10)$  at age ten, if his slower rate of growth is held. His M.A. at age ten is eight years. In both of these cases, the I.Q. is kept constant by the fact that *one* year of mental age at five corresponds to *two* years of mental age at ten. The I.Q. of any child, superior, normal, or inferior, will remain constant over the age-interval spanned by the tests, as long as test scores exhibit *increasing spread* at succeeding ages, and growth is represented by *straight lines*.

Figure 2 shows another hypothetical mental growth curve which will yield a constant I.Q. The first important characteristic to note here is the steady decrease in the unit of mental measurement as chronological age increases. The second is the logarithmic shape of the growth curve. As in Figure 1, the middle curve pictures normal or average growth in mental ability, the upper and lower curves, superior and inferior ability, respectively. All three curves exhibit negative acceleration, *i.e.*, bend in more and more toward the base line, the middle curve becoming parallel to the base line at approximately fourteen years. It will be noted that the spread (SD) of the test scores at succeeding ages is constant, as shown by the distance between the curves measured by lines perpendicular to the base line. The increasing spread of test scores shown in Figure 1 is offset in Figure 2 by the *decrease* in the mental unit of measurement (*i.e.*, the M.A.); and it is this steady and regular decrease in the mental growth unit with the subsequent bending-in of the growth lines which gives the constant I.Q. As before, a child aged five who tests six on the scale, with an I.Q. of 120, will at ten years of age test twelve.

His I.Q. remains constant at 120, provided his original rate of growth is maintained. In like manner, a child whose I.Q. at age five is 80 (M.A. four) will at age ten have an I.Q. of 80 (M.A. eight), provided his slower rate of growth remains constant.

### 3. The Mental Growth Curve of the Stanford-Binet

If we consider the Stanford-Binet Scale in the light of the previous discussion, it will be evident that its data must follow the type of growth curve shown in Figure 1. Binet M.A.'s of six, seven, and eight years, it must be remembered, simply represent the average scores made by unselected six-, seven-, and eight-year-old children. If the average seven-year-old child scores 35 on a test, and the average eight-year-old 42, a score of 35 represents an M.A. of seven years, and a score of 42 an M.A. of eight years. A year of mental age, therefore, must correspond to a year of chronological age, because of the way in which the age-scale is constructed; and if we plot Binet M.A.'s against C.A.'s, the relationship will *necessarily* be linear (straight line), as in Figure 1. The spread of scores in the Stanford-Binet Scale at succeeding ages (in terms of the M.A. as unit) becomes *steadily greater* as we go up the age-scale. Terman (64), for example, found that the middle 50 per cent. of his six-year-olds fell within a range of ten mental months around the median child of that age; the middle 50 per cent. of ten-year-olds within a range of sixteen mental months around the median ten-year-old child; and the middle 50 per cent. of fourteen-year-olds within a range of twenty-six mental months around the median fourteen-year-old child.

The I.Q. of a child of nine, therefore, is comparable to the I.Q. of a child of six or ten, because the age-progress curve of the Stanford Revision is a straight line with diverging inferior and superior growth lines, as shown in Figure 1. The I.Q. from such a scale will remain constant over the interval to which the scale applies. In terms of I.Q., as measured by Stanford-Binet, the middle 50 per cent. of the children at each age between four and fourteen fall roughly within eight to ten points above and below the median I.Q. of 100. Given two children, both of whom have I.Q.'s of 110, each is in the same degree superior to the median child of his own age group.

It cannot be too much emphasized that the I.Q. of the Stanford-Binet will remain constant *only* over the age range to which the

scale is applicable. As we have pointed out, the M.A. is an arbitrary unit, being simply the C.A. corresponding to an average score. If the average scores of successive age groups cease to show progressive increases, the M.A. ceases to be a useful measure, since it no longer serves to differentiate one age group from another. When a scale records the M.A.'s of eleven-, twelve-, thirteen-, and fourteen-year-old children as being, in each case, twelve years, say, it no longer separates one age group from another in terms of performance. Moreover, when successive age groups no longer show increased scores, the mental growth curve can no longer remain a straight line, but must bend in toward the X-axis or base line. This is actually what happens in the case of the Stanford-Binet Scale. Logically, it seems unduly optimistic to expect mental growth to progress indefinitely along a straight line. Data employed by Terman in standardizing the upper levels of the Stanford-Binet Scale support this contention experimentally. Terman (64) reports the mean M.A. of thirty-two high-school students, with a mean C.A. of eighteen years, to be  $16 \pm .67$  year on the Stanford-Binet Scale; and the mean M.A. of 180 adults, tested by Knollin (33), to be  $14.1 \pm 1.38$  years.

These results show clearly that the Stanford-Binet ceases to differentiate in terms of M.A. somewhere between fourteen and eighteen years. Hence, an adult of thirty or forty will perform no better on the scale, ordinarily, than a child of fourteen or sixteen. Basing his conclusion on all of his available data, Terman set sixteen years as the point where the average person is "mentally mature" according to the Stanford-Binet Scale. This means that the curve of mental growth for the Stanford-Binet is no longer a straight line beyond age sixteen, but tends to become parallel to the X-axis close to this point.

It is sometimes not clearly understood that the failure of the Binet and its derivatives to differentiate in terms of M.A. beyond the early 'teens makes it impossible for the I.Q.'s of superior children to remain constant. Mental age, however, inevitably lags farther and farther behind chronological age as the superior child grows older. The maximum mental age which an adult can earn on the Stanford-Binet Scale is nineteen and one-half years. (See p. 12.) Hence, the maximum I.Q. which an adult can achieve, no matter how bright he may be, is  $19.5/16$ , or 122 (p. 9). If a child of four has a mental age of eight, his I.Q. is 200. Since the maximum mental age which

this child can earn is 19.5, his I.Q. will remain constant, assuming mental growth to continue at the same rate, until he is nine years nine months old, when his I.Q. is  $19.5/9.75$ , or 200. After he is ten years old, the I.Q. of this child must necessarily decrease below 200, becoming 122 when he is an adult. This drop in I.Q. is inherent in the age-scale, and applies only to superior children. With normal and retarded children the scale does not show this inadequacy, since the M.A. in such cases is never greater than the C.A.

Let us summarize briefly the results of this rather long discussion.

(a) The I.Q. of the Binet tests and their revisions is theoretically constant up to the early 'teens because the standardization of the age-scale results in a straight-line growth curve, with diverging inferior and superior lines of growth, as pictured in Figure 1. (b) From about age fourteen on, however, there is no longer a corresponding increase in Binet M.A. with increase in C.A. At this point the ratio M.A./C.A., or I.Q., is no longer a valid measure of mental capacity, and the growth curve becomes parallel to the base line. The failure of the age-scale to record further increases after fourteen to sixteen years may be owing to the fact that mental growth in the functions measured has actually reached a maximum. A more probable explanation is that simple verbal tests are no longer capable of measuring increases after the early 'teens, having reached here the limit of their ability to differentiate one individual from another. (c) The I.Q. technique should not be employed with superior children in their 'teens, nor ordinarily with normal or superior adults.

#### 4. The Curve of Mental Growth When Measured in Equal Units

As pointed out in the preceding section, mental growth as pictured by the Binet tests cannot be represented by a straight line after the early 'teens (fourteen to sixteen years). This finding suggests that the progress of mental growth with age, if measured in terms of a unit more stable than M.A., might be more truly represented by such a curve as is shown in Figure 2. Probably the most nearly "true" picture of mental growth as measured by such tests as the Binet is shown in Figure 3. This curve was drawn by Thurstone (69) from a mathematically determined absolute zero, the successive increments of mental growth from year to year being measured in terms of a constant achievement unit. The data, upon which the curve is based, consist of records from more than 3,000 normal and defective Lon-

don school children measured by Burt (6) upon an adaptation of the Binet scale. Thorndike's curve in Figure 4 (68), which is based upon results from his CAVD Examination (p. 36), also depicts the relationship of age to general intelligence when intelligence is measured in equal units. These two growth curves are far more authentic than the ordinary curve of performance with age, drawn from some arbitrary starting point (say three years), and plotted in terms of a growth unit (such as items done correctly or points earned) which varies from year to year. Note that both curves are parabolic in in

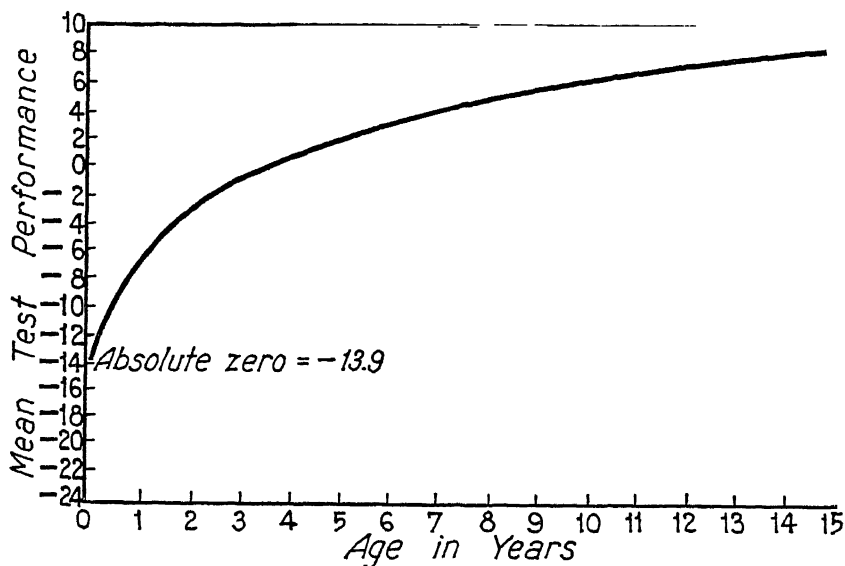


Figure 3.—MENTAL GROWTH ON THE BINET TESTS MEASURED IN EQUAL UNITS  
(THURSTONE [69])

form. Thurstone's curve rises rapidly from 0, and then more and more slowly to become parallel to the base line at fourteen to sixteen years. Thorndike's curve also rises rapidly at first, but does not reach its level until close to twenty years, owing chiefly to the fact that the CAVD Examination is able to measure changes in the upper age levels (p. 37), where the Binet revisions do not differentiate.

It is evident from Figure 3 that a mental year becomes a progressively smaller and smaller increment as chronological age increases. This means that a child of five who is one year accelerated in mental age is as far ahead of his age group as is a child of ten

who is two years accelerated. These relations are clearly indicated in Figure 2.

### 5. Experimental Evidence for the Constancy of the I.Q.

So many experimental studies (71) have been made of the constancy of the I.Q. as shown by retests, that we can cite only a few of the more extensive here. Terman (61) studied the constancy of the I.Q. for 315 children, retested at intervals of from one to seven years. The majority were retested after an interval of more than one year, about one-third being measured after five years or more.

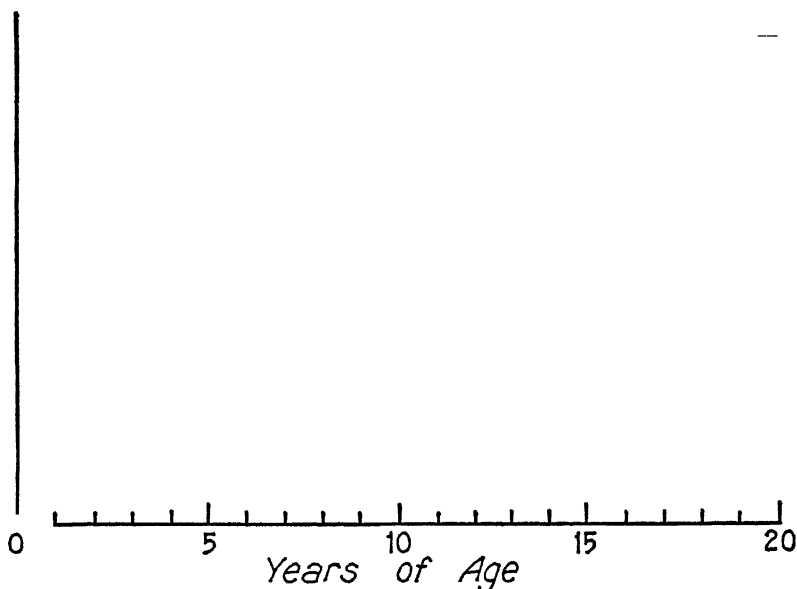


Figure 4.—MENTAL GROWTH ON THORNDIKE'S CAVD INTELLIGENCE EXAMINATION  
MEASURED IN EQUAL UNITS (THORNDIKE [68])

Terman's conclusions were (a) that the central tendency of changes in I.Q. is represented by an increase of 1.7; (b) that the middle 50 per cent. of changes falls between  $-3.3$  and  $5.7$ ; and (c) that the P.E. of a prediction based upon a single test is about 4.5 I.Q. points. The reliability coefficient of the Stanford-Binet Scale was .93.

Hildreth (30) has substantiated Terman's findings, in large part, upon a population of 441 school children. Changes in Stanford-Binet I.Q. were studied in successive retests. The number of times the test was given to each child varied from two to eight, the children's ages ranged from three to eighteen, and their I.Q.'s ranged



from 80 to 185. The median change in I.Q. was .96 of a point, the middle 50 per cent. of changes being from  $-3.5$  to  $5.7$  points. The correlation between all pairs of tests was  $.81 \pm .01$ .

P. Cattell (9), in connection with the Harvard University Growth Study, has reported the results of retests made on 1,183 children after intervals of from three months to six years. These findings are given in Table I. It appears that the short intervals (three to six months) give on the average an increase of from four to five points in I.Q., a change which may be attributed largely to practice. The average change after six months is negligible. Cattell found a definite tendency for young children of high intelligence to gain, and those of low intelligence to lose slightly in I.Q. as they grow older. The tendency of the I.Q. to decrease with age in the case of feeble-minded children was reported earlier by Kuhlmann (38) as a result of retests on 639 children over a period of ten years.

To summarize briefly, experimental studies indicate that the I.Q.'s of normal children remain substantially constant from year to year. There is a tendency, however, for bright young children to gain and dull children to lose slightly in I.Q. with age increases. This does not contradict our earlier statement (p. 18) that the I.Q.'s of bright children tend to decrease, at the upper-age levels, owing to the nature of the scale.

TABLE I  
MEDIAN BINET I.Q. CHANGES AFTER VARYING INTERVALS OF TIME  
(from P. Cattell [9])

Months between Tests	No. of Cases	Median Differences	P.E.
0-3 .....	18	+5 0	$\pm 1.1$
3-6 .....	54	+3 8	$\pm 0.8$
6-12 .....	308	+0 2	$\pm 0.4$
12-18 .....	174	+2 0	$\pm 0.6$
18-24 .....	64	-0 5	$\pm 0.9$
24-36 .....	92	-0 2	$\pm 0.9$
36-48 .....	293	-2 7	$\pm 0.4$
48-60 .....	51	+1 0	$\pm 1.5$
60-72 .....	329	-0 1	$\pm 0.4$
0-72 .....	1383*	-0 02	$\pm 0.22$

\* Some children were used in more than one comparison

## 6. Factors Affecting the Constancy of the I.Q.

Authorities are not fully agreed as to the influence of variations in physical growth upon measures of general intelligence. Among the many factors which may affect the constancy of the I.Q., we

may list as obviously important (a) irregularities in physical and mental growth; (b) physical defects, diseases, *etc.*; (c) practice and coaching; and (d) environmental influences. Dearborn (16) and Gesell (25) have both reported cases of children whose physical and mental growth has apparently slowed down markedly, and later shown a sudden spurt; and children in whom a slowing down of mental growth has not been later repaired. Such cases as these are associated usually with decided physical and nervous symptoms (6). It is obvious that severe sensory handicaps, such as deafness and blindness, will affect intelligence test results directly, but the ordinary diseases and afflictions of childhood seem to have little deleterious effects. M. C. Rogers (50) reported no significant improvement in the I.Q.'s of twenty-eight children retested six months after the removal of diseased tonsils. Even twelve months after the operation there was little change in I.Q. Fox (20) in a study of twenty-two children, mostly thyroid cases, found little effect upon I.Q. of glandular therapy, the net gain after treatment being 1.5 points.

Graves (28) has made a very thorough study of the effects of coaching upon Stanford-Binet I.Q. Definite coaching upon the tests themselves gave decided increases of as much as thirty-two months in mental age. After twelve months the effects of coaching still persisted, the coached group being slightly superior to the control group. Since intelligence test booklets have been widely disseminated, examiners must guard against the possibility that a child has been definitely prepared for an intelligence test.

The effects of environmental influences upon general intelligence tests are so wide and varied that they can be indicated here in a brief way only. In a study of the effects of language handicaps upon intelligence tests, Mead (42) administered the Otis Test of general intelligence to 276 Italian and 160 American school children in Grades 6 to 10 in a small New Jersey town. On the basis of the amount of English spoken by their parents in the home, the Italian children were divided into the following four groups: only Italian spoken at home; mostly Italian with some English; mostly English with some Italian; and only English. The mean test scores of these four groups were 65, 70, 74, and 81—a steadily increasing mean score with the amount of English spoken. The Italians were closest to the American score in the Arithmetic Test, and farthest removed

in the Proverbs Test. These results indicate the importance of comparable language background in comparing groups or individuals.

Klineberg (35) has shown the necessity of considering differences in culture, training, and temperament when evaluating mental test performance. Burks (5), and Freeman (21), whose studies are reported in some detail in Chapter V (p. 189), have demonstrated the effect of home environment upon intelligence test ratings. Whenever children have had much the same social, educational, and recreational opportunities, one is usually justified in classifying them on the basis of mental test ratings as retarded, normal, or bright in verbal or abstract intelligence. But the examiner must always remember that comparisons are permissible only when environmental differences are absent, or at least negligible.

### 7. Mental Maturity as Measured by the Age-scale

Kuhlmann located the M.A. for the average adult at fifteen years on his scale, while Terman located the average adult level at sixteen years and the superior adult level at eighteen years on the Stanford Revision (64). In order to find the I.Q. of an adult (*i.e.*, to compare his brightness with that of the normal) we divide the mental age earned on the scale by fifteen, if the Kuhlmann, and by sixteen if the Stanford-Binet is used, whether the individual is twenty-five, forty, or sixty years of age. A man of forty who earned an M.A. of ten years on the Stanford-Binet would have an I.Q. of  $10/16$ , or 63, and would, accordingly, be feeble-minded.

Other psychologists have placed the normal adult level somewhat lower than fifteen or sixteen years. Thus Pintner (45) puts it at fourteen years, and Dearborn (16) at fourteen and one-half years, these lower levels being more in accord with the Army Alpha results (p. 34). In setting the point at which general intelligence "matures" at a given age, psychologists do not imply that *all* mental or intellectual growth stops at this point. All that is meant is that for most adults, the particular scale employed ceases to record progress beyond the given age level.

One must carefully distinguish between normal adult level and the limits of mental growth on an age-scale. The failure of the Binet scale and its revisions to measure mental progress beyond the early 'teens is—at least in superior individuals—owing to a lack of discriminating tests at the upper levels rather than to a marked slowing

down in mental power. When adequate scales are available, the evidence is that at least the ability to learn increases up to the early twenties or beyond (65).

In a study of more than 8,000 high-school students in Grades 9 to 11, Thorndike (67) found steady score increases from ages fourteen to eighteen upon an extensive battery of mental tests. Teagarden (59), who tested 408 children twelve and one-half to twenty years old with Stanford-Binet and several group tests, found increases on Stanford-Binet up to eighteen years. It should be noted that the upper-age groups were small, however, and probably selected, and that the test gains were slight. Increases in score appear upon CAVD up to about twenty years of age (68).

The results just quoted are based upon the performance of individuals of above-average ability. In less highly selected groups there is little, if any, increase in general intelligence test score with age, above fourteen to sixteen years. This is shown in an extensive study of change in mental test score with age, conducted by H. E. Jones (45). Jones' sample embraced 1,151 individuals from eleven to fifty-four years of age, all resident in a rural New England community. It represented a good cross section of native white American stock. Median scores on Alpha rose steadily up to fourteen years. From fourteen to sixteen the gain was twelve points; but after sixteen the medians remained at the same level, with minor fluctuations, up to about forty years.

High-school and college students, who have already been selected for abstract ability, will continue to show increases upon an adequate scale beyond fourteen or even beyond sixteen years. But the large majority of the general population, 40 per cent. of whom drop out of school by the eighth grade, is more fairly represented by an adult level of fourteen to sixteen years on verbal or abstract tests of general intelligence.

### 8. Distribution of I.Q.'s in the General Population

The distribution of I.Q.'s for 905 unselected children on the Stanford Revision (60) is shown in Figure 5. This distribution follows closely the normal or probability distribution. The various degrees of intelligence as given by the scale are grouped around 100 I.Q. as the central point, decreasing gradually toward the upper and lower extremes. As there are no sharp lines of division from one

group to another, the classification of children into normal, superior, and inferior groups is an arbitrary matter and if strictly adhered to may lead to injustice. Certain broad divisions, however, have been made by psychologists, and these may be followed as a general guide. The number and percentage of children in the general population which may be expected to fall within each classification are given in Table II. It will be noted that children who test below 70 I.Q. are

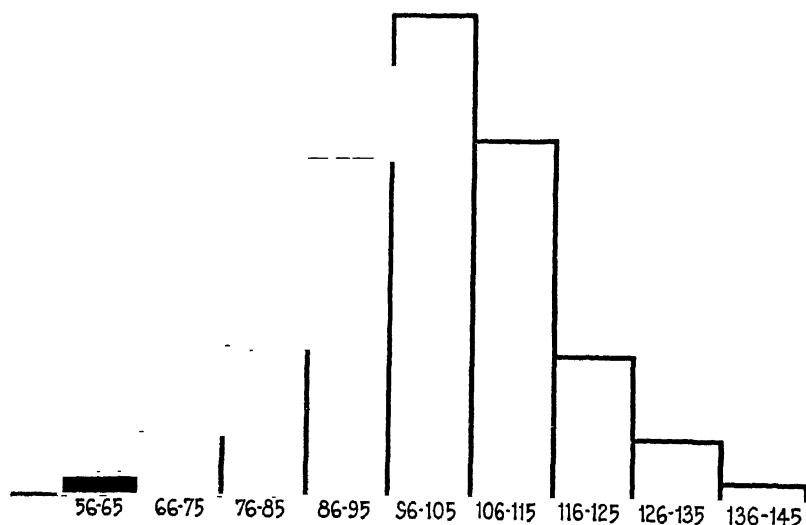


Figure 5.—DISTRIBUTION OF I.Q.'s (STANFORD-BINET) IN THE GENERAL POPULATION

classified as feeble-minded. The percentage of children in this or any other category must not be thought of as constant, however, for it depends directly upon the selection and general character of the group tested.

TABLE II  
APPROXIMATE DISTRIBUTION OF I.Q.'s IN THE GENERAL POPULATION  
(from Sandiford [53])

Classification	I.Q.	Percentage of All Children Included
"Near" genius or genius. . . . .	Above 140	.25
Very superior . . . . .	120-140	6 75
Superior . . . . .	110-120	13 00
Normal, or average . . . . .	90-110	60 00
Dull, rarely reported as feeble-minded . . . . .	80-90	13 00
Border-line, sometimes dull, often feeble-minded . . . . .	70-80	6 00
Feeble-minded . . . . .	Below 70	1 00

### 9. Value of the I.Q. in School Work

The correlation of Binet I.Q. and school achievement has been found to vary from about .40 to .75, the exact degree of relationship depending upon the size of the group and the ages of the children measured (45). One reason for the substantial correlation between I.Q. and school achievement may be found in the way in which the age-scale was validated. In constructing the Stanford-Binet Scale, Terman used school success as one check upon the validity of his tests (p. 10). Terman (64) obtained a correlation of .45, for instance, between teachers' ratings for school work and the I.Q.'s of 504 children, ranging from below 80 I.Q. to above 120 I.Q. Many later investigations have been made of the correlation between Binet I.Q. and school work. Of these studies we shall cite a few representative ones in which the correlations were based upon samples large enough to give trustworthy results. Dickson (17) has reported a correlation of .73 between ratings for school work and the mental ages of 149 first-grade children. Proctor (48) obtained a correlation of .55 between Stanford-Binet I.Q. and school marks in a group of 102 high-school pupils. Burt (6), using an adaptation of the original Binet Scale, obtained a correlation of .68 (age variability held constant) between Binet M.A. and school achievement in a group of 300 London school children, seven to fourteen years old. Witty and Taylor (72) obtained an  $r$  of .63 between Binet M.A. and educational achievement in a group of 522 children in Grades 4 to 6. Freeman, *et al.* (21), report an  $r$  of .70 between a school-achievement index and Stanford-Binet I.Q. in a group of 348 children.

M. V. Cobb (11) has made an extensive study of the relationship existing between I.Q. and school marks in a group of 1,016 public school children. School marks were reported as A, B, C, D, and E, denoting respectively excellent, good, passing, conditioned, and failed. The correlation (which was calculated by the writers) between marks and I.Q. is .74. Only four children below 70 I.Q. made passing grades, while only two above 90 I.Q. failed outright.

The correlation between school achievement, as measured by standard educational tests, and general intelligence, as measured by group tests, is usually even higher than the correlation between school marks and individual intelligence tests. Kelley (33) has shown that the estimated true correlation between general intelligence as measured by the N.I.T. and school achievement as measured by

the Thorndike-McCall Reading Test and the Woody-McCall Arithmetic Test was .93 for 200 children in Grades 4 to 8. Other correlations cited by the same author between tests of general intelligence and the Stanford Achievement Test show equally high relationships. These latter results are especially valuable as showing the "community of function" in general intelligence tests and measures of school achievement, because standard educational tests are far more reliable than school grades (p. 58).

The real reason, of course, for the marked relationship between school achievement and I.Q. is that both tap the same abilities. As pointed out above, the Binet Test and its revisions are concerned primarily with the measurement of abstract or verbal ability, especially in the ages above seven years. Tests of vocabulary and word meaning, as well as tests of the comprehension of language and language relationships, play an important part both in school work and in determining a child's I.Q. Terman remarks that the vocabulary test has "far higher value than any other single test on the scale" (60). He states further that "our statistics show that in a large majority of cases the vocabulary test alone will give us an intelligence quotient within 10 per cent. of that secured by the entire scale." A correlation of .91 between the Stanford-Binet Vocabulary Test and Stanford-Binet M.A. has been reported by Terman in a group of 631 children, Grade 1 to first-year high school (62). This  $r$  is "inflated" because of the wide age range (M.A.'s range from five and one-half to nineteen years) and of the fact that the vocabulary test is a component part of Stanford-Binet M.A. But the correlation would doubtless still be very high in a more homogeneous group, as shown by  $r = .82$  between vocabulary and Stanford-Binet I.Q. obtained by Burks (5) in a group of eighty-seven children in which the age-variability factor was controlled.

The I.Q. is of value in enabling one to discover in a relatively short time the probable upper limit of scholastic attainment for a given child. Terman (63) has drawn up the following table (Table III) which gives the mental age "normally" required to do the work of the different school grades. Children in grades corresponding to their mental ages will usually be found doing work of average quality, while children in grades above their M.A. will experience increasing difficulty. If a child's I.Q. is 80, for instance, his maximum mental

TABLE III  
MENTAL AGE STANDARDS FOR THE DIFFERENT GRADES  
(from Terman, *et al.* [63])

Grade	Standard Mental Age	Mental Age at Mid-Grade
I	6-6 to 7-5	7 years
II	7-6 to 8-5	8 years
III	8-6 to 9-5	9 years
IV	9-6 to 10-5	10 years
V	10-6 to 11-5	11 years
VI	11-6 to 12-5	12 years
VII	12-6 to 13-5	13 years
VIII	13-6 to 14-5	14 years
1st Yr. H. S.	14-6 to 15-5	15 years

age on the Stanford-Binet will be approximately thirteen years. This means that such a child cannot be expected to do school work successfully beyond the seventh grade.

We may set down the educational expectation of children with different I.Q.'s as follows:

(a) The average child of 60-65 I.Q. remains in the first grade until he is ten or eleven, reaching the fifth grade when he is about fourteen or fifteen (his mental age will then be nine years). Although this child's mental development is inadequate for fifth-grade work (see Table III), because of his advanced chronological age he is usually pushed ahead after repeating a grade two or three times. Children who are considerably over-age for their grade will usually be found to have been promoted on the basis of length of service rather than accomplishment. The child of 60-65 I.Q. is rarely found above the fifth grade.

(b) The average child of 75-79 I.Q. reaches the fifth grade by the time he is thirteen and the eighth grade by the time he is sixteen or seventeen. Ordinarily his education stops here.

(c) The average child of 80-84 I.Q. spends two years in the first grade and completes the eighth grade, if at all, two or three years behind schedule. Rarely does he go beyond this point.

(d) The average child of 85-95 I.Q. will ordinarily be about one year over-age for his grade. Entrance to high school is practically barred for children under 90 I.Q. If the school is small, however, so that much individual attention can be given to the child, and if the standards are not too rigid, a child of 85-95 I.Q. may graduate from high school by dint of much persistence and the repetition of many subjects.

(e) Children with I.Q.'s from 95 to 105 constitute about 40 per



cent. of the elementary school population. If not retarded by illness or loss of schooling through various other causes, these children will complete elementary and high school. Even when exceptionally industrious, however, they are not good college material, and probably should not be encouraged to seek education beyond the high school. In a survey of more than 6,000 high-school seniors in Indiana, Book (3) found that almost as many students in the lower intelligence levels were intending to go to college as in the upper levels. Many of the brightest children were not preparing for college. The great need for educational advice and guidance here is obvious.

(f) Children with I.Q.'s of 120 and above are usually one to two years accelerated and reach the eighth grade by the time they are twelve or thirteen years old. Above 130 I.Q. acceleration may be even greater. All too often, however, the superior child is kept in a grade normal for his chronological age because the teacher or principal thinks he is too young to be promoted, or because the machinery of the school is not adapted to extra promotion.

As individuals who are over sixteen years of age are not accurately measured by an age-scale, the I.Q. is of little value in the upper high-school grades and at the college level. Such tests as have been made indicate that the I.Q. of the average high-school graduate is probably close to 110, and that a student with an I.Q. below 90 has little chance of graduating from a first-class high school. Unless a student's I.Q. is 115 or more it is probably unwise to advise him to go to college (47).

The intelligence test does not presume to give information beyond the prospective abilities of the child on the abstract or verbal level. But, no matter what the nature of the problem, the I.Q. does limit the field of inquiry. Given two children both failing in school, one would look for very different causes if the first child had an I.Q. of 120 and the second an I.Q. of 85. In every case, to be sure, knowledge of the I.Q. should be supplemented by a study of the child's personality, home conditions, outside interests, *etc.*, if one is to advise him intelligently. In the hands of a skilled clinical psychologist, the mental examination will often reveal much concerning the temperament, home training, attitudes and habits of work of a child. There are many tests now available which will prove useful in a study of special abilities, dominant interests and temperamental and character traits. (See Chapters II and III.)

It may be helpful to give several concrete instances of how intelligence tests can be used in the study of definite educational problems (17).

Harriet S. entered the low first grade at age 6-1; M.A. 5-8; I.Q. 93. Her school progress showed acceleration until she was in the high fifth grade four years after her school entrance. Here is a case where a child of apparently near normal capacity has made rapid progress. Why? Her teachers' explanation is that she is an "extremely hard worker." When we add to this the fact that her scholarship record for five consecutive terms was only a "Passed" and also the teacher's statement in the high fifth grade that "she is beyond her depth now," we have evidence of the fact that hard work cannot take the place of innate mental capacity. Far better for Harriet if she had been allowed to proceed at the normal rate of progress as indicated by her mental test. She would have been saved the humiliation of getting to the place of discouragement where she was clearly "beyond her depth."

Henry (a high-school boy) wished to become a civil engineer. He was an unusually conscientious, hard-working, over-age boy of seventeen years, in the second semester of the ninth grade, repeating algebra because his teacher told him that "he must have a good foundation in algebra in order to do satisfactory work in the later mathematics required for engineering." The pathos of the situation becomes striking when we know that this boy, who spent fruitless hours of labor upon complicated problems in algebra, had an I.Q. of 83, that he had a record of repeated failure in the upper elementary school, with marks of "good" in conduct and "good" in attendance, and that in the first semester in high school he had earned barely passing marks in two subjects and failure in two others. The father, a rather shiftless carpenter, "didn't care if the boy went to school." The mother, a quiet, persistent woman, was determined that Henry, who is her youngest and favorite son, "should get a good education and hold a good job," for his two elder brothers are now earning a bare living at hard menial labor "because they didn't like school and quit, one in the fifth, the other in the sixth grade."

This case is illustrative of a dogged persistence not altogether uncommon in high-school pupils in which effort that might otherwise be turned to good account wastes itself on impossible goals. There is not one chance in 100 that Henry will finish three years of standard high-school work. Why was he advised to repeat algebra "in order to do satisfactory work in the later mathematics required for engineering"? Apparently because his mathematics teacher knew more and thought more about the requirements for civil engineering than he knew or thought about the characteristics of the boy. . . . This boy and his parents need counsel as to an educational objective more clearly within the range of probable attainment. . . .

Frances S. entered the low first grade in January, 1918. Chronological age 6-0; M.A. 8-8; I.Q. 144. In May of the same year she was promoted to the

high first grade, and in the next semester completed the work of that grade as well as the requirements of the low second. In January, 1922, she had completed the high fifth with excellent scholarship. Thus she accomplished five years' work in four years. Her chronological age during the term in the high fifth grade was ten years, while her mental age was above fourteen. The case illustrates the fact that a superior I.Q. furnishes a reliable basis for predicting accelerated progress, but that such progress all too frequently does not keep pace with the mental development of the child. There develops an increasing degree of mental over-ageness in the superior child who is given inadequate opportunity for acceleration.

In concluding this section, it cannot be too much emphasized that many failures both in elementary and high schools occur among children whose general intelligence is adequate to do the work of their grade. Social and temperamental factors, interest, will-to-do, home-training, lack of stimulating surroundings, all contribute to many of these failures. Health may also be an important consideration (32).

#### GROUP TESTS OF GENERAL INTELLIGENCE

As we have already pointed out (p. 5), the group test of general intelligence is in form much like the ordinary school examination. Instead of demanding specific school information, however, its aim is to gauge the individual's ability to follow fairly complicated directions, read accurately and understandingly, solve problems, perceive relations represented by language or numerical symbols, and use previously acquired information. The tests most frequently included in a group test battery are ordinary mental arithmetic problems, opposites, analogies, range of common information, and tests involving logical selection and classification (12). The group test of general intelligence attempts to discover potential learning ability by measuring how much one has already learned. It is thus a gauge of mental alertness as well as an estimate of acquired information.

#### THE ARMY ALPHA INTELLIGENCE EXAMINATION

The group test of general intelligence grew out of the desirability of testing large numbers of children or adults at the same time, without resorting to elaborate techniques or apparatus and without requiring much training on the part of the examiner. The first urgent need for a group test of general ability arose during the World War when it became necessary to sift out from the draft, men unfit for military service through lack of intelligence; to select the more

intelligent men for further training, or for special service; and to provide more nearly balanced companies and regiments. The Army Alpha Test for literates and the Army Beta Test for illiterates were devised to meet this situation. These tests were administered to more than 1,750,000 men between September, 1917, and January, 1919. Because of the extensive data collected, Alpha results give the best picture of the relative abilities upon such an examination of various occupational, racial, and other groups, drawn from the literate male population of the United States. Furthermore, since many of the group tests which have been devised since Alpha follow it closely both in content and in method of construction, results from this examination are worth examining at greater length.

So much has been written about the Army Alpha Test itself (78) that only a brief description is necessary here. Army Alpha was validated against various criteria: (1) officers' ratings of their men; (2) Stanford-Binet M.A.'s; (3) the Trabue Language Completion Scales; (4) school marks and (5) teachers' ratings for intelligence. Alpha's correlations with these various criteria ranged from .50 to .90. The reliability coefficient of Alpha is about .95 for unselected male groups, the P.E. of an Alpha score being about five points. There were five forms of Alpha, each form containing eight sub-tests. These component tests were (1) Following Directions; (2) Arithmetic Problems; (3) Practical or Common Sense Judgment; (4) Synonym-Antonym; (5) Disarranged Sentences; (6) Number Series Completion; (7) Analogies; (8) Information. The items in each sub-test were arranged in order of difficulty, and a time limit set for each sub-test. The maximum score which can be obtained on the Army Alpha Test as a whole is 212.

### 1. Distribution of Alpha Scores in the General Population

The distribution of general intelligence (as measured by Alpha) in the general male population of white, native-born, enlisted men is shown in Figure 6. This graph is based upon the records of 51,620 enlisted men, selected to give a representative sample of the American male population. The average Alpha score made by native-born white enlisted men was 59; by foreign-born white enlisted men 47; and by white officers, 139 (43). In order to enable a broad classification of soldiers into ability groups, superior, average, inferior, and the like, letter grades were assigned to certain score intervals on

the scale. These letter grades were based upon the actual Alpha score distributions of large representative groups. Table IV gives the score ranges corresponding to each letter grade and the percentage of enlisted men receiving each grade. Note that the average

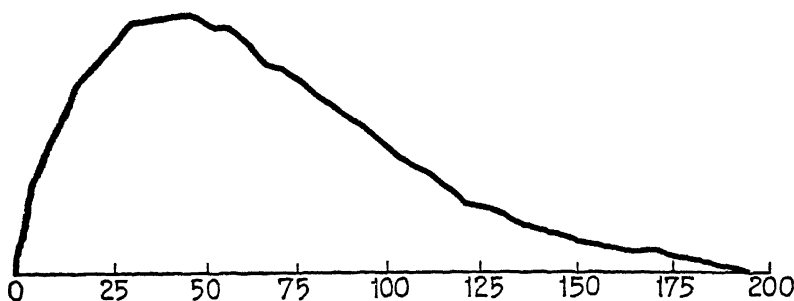


Figure 6.—DISTRIBUTION OF GENERAL INTELLIGENCE (ARMY ALPHA SCORES) FOR WHITE, NATIVE-BORN, ENLISTED MEN

officer received a rating of A on the basis of his score, the average white enlisted man a rating of C.

TABLE IV

PERCENTAGE OF LETTER GRADES MADE ON ALPHA ( $N = 93,973$ )

(from *Memoirs*, National Academy of Sciences [13])

Letter Rating	A	B	C+	C	C-	D	D-
Range of scores . . . . .	212-135	131-105	101-75	74-45	41-25	24-15	11-0
Per cent. of white draft receiving grade . . . . .	4 1	8 0	15 0	25 0	23 8	17 1	7 0

## 2. Binet Mental Age and Alpha Scores

In order to evaluate Alpha scores in terms of Stanford-Binet mental age, Alpha scores made by a group of 653 men were transmuted into mental ages earned by these same men upon the Stanford-Binet Scale. Binet M.A.'s corresponding to Alpha letter grades are shown in Table V. The average score of white enlisted men (*i.e.*, 59) corresponds to a Stanford-Binet M.A. of about thirteen years. When the officers were combined pro rata with the enlisted men, the average "adult level" was approximately fourteen years in terms of Stanford-Binet M.A. That is, the average adult level on Army Alpha was two years below the adult level (*i.e.*, sixteen) of the Stanford-Binet.

This discrepancy between the two tests is not very disturbing when one examines (a) the tests themselves, and (b) the method

TABLE V

STANFORD-BINET MENTAL AGES CORRESPONDING TO THE VARIOUS ALPHA LETTER GRADES  
( $N = 653$ )

(from *Memoirs*, National Academy of Sciences [43])

Letter Rating	A	B	C+	C	C-	D	D-
Mental ages . . . . .	18+	17 9-16 5	16 4-15	14 9-13	12 9-11	10 9-9 5	9 4

whereby norms were obtained upon each. Alpha and Stanford-Binet, although containing much in common, are clearly not identical either in content or in method of administration. The correlation between the two tests was .81 in the sample of 653 men. The average adult level of the Stanford-Binet was based upon results obtained with about 400 adults, many of whom (high-school students and business men) were obviously selected with reference to the general population. On the other hand, Army Alpha's mean for the general population was obtained from more than 100,000 men drawn from all walks of life. The M.A. of the general population in terms of Stanford-Binet is, therefore, closer to fourteen than to sixteen years.

The above conclusion has often been taken to mean that the average adult in the United States has the mind of a fourteen-year-old child. Such an interpretation is, of course, absurd. An average adult level of fourteen on the Army Alpha Test simply means that the average soldier did about as well on the test as the average first-year high-school student. This is far from a discreditable showing, when one considers that many of the men who took Alpha possessed meager schooling, and that often not very recent. The effect upon a verbal or language examination of continued practice in reading and in arithmetic is shown clearly by the fact that bookkeepers made an average score of 101 on Alpha, as against the average score of 63 made by the general machinist. It seems hardly probable that this difference in score arose entirely from differences in native capacity. Part of it, at least, can be attributed to the more habitual use by the clerical worker of the kind of operations called for by the Alpha examination.

### 3. Revisions and Present Use of the Army Alpha

Army Alpha was used extensively after the close of the war, in high schools and colleges as well as in business institutions. In recent years it has been largely superseded in schools by tests more suitable in content and of greater difficulty range. However, because of the large number of occupations for which there are Alpha norms, and the wide variety of the other groups to which it has been given,

Alpha is still employed extensively as a test for adults of moderate, i.e., common-school, education.

In 1925 the Psychological Corporation prepared a revision of Army Alpha (p. 41). This is a compilation of the best items from the five original forms, leaves out the more "military" questions, and is better adapted for use with ordinary civilian groups. Bregman (4) has drawn up a percentile table, based upon the Army results from which an individual's score can be converted into a twenty-fifth, sixtieth, or eightieth percentile rank in the general population. A still more recent revision of Army Alpha prepared by F. L. Wells was published in 1932 by the Psychological Corporation under the title, *Revised Alpha Examination, Form 5*. In this revision Test 1 is replaced by an addition test, and Test 8 by a verbal directions test. The recording of answers and scoring of several tests has been simplified. The directions card accompanying the test states that the revision has been for the purpose of making the test easier to give and to score; more appropriate for educational and business use; and to keep its contents from going out of date. A table of percentile equivalents to test scores is printed on the directions card. This adds much to the practical value of the test.

Table VI gives the medians and range of the middle 50 per cent. of scores made by men of various occupations who were enlisted in the army. This table shows clearly the progressive increase in Alpha score from the unskilled laborer level through the business to the professional groups. It enables an executive to estimate the vocational possibilities (as far as verbal or scholastic intelligence is concerned) of a man who scores 60, say, or 100 on Alpha.

#### THE CAVD INTELLIGENCE EXAMINATION

The CAVD Intelligence Scale (68) has been selected for special mention because it is superior in construction to other tests of general intelligence. CAVD is composed of tasks involving the completion of sentences (C), arithmetic problems (A), vocabulary (V), and directions (D), the last named often requiring comprehension of sentences and paragraphs. Forty items in all—ten each of C, A, V and D—are placed at seventeen levels of difficulty ranging progressively from level A to level Q; that is, from a composite (level A) at which 50 per cent. of the forty items can be solved correctly by an individual of M.A. three years, to a composite (level Q) at

TABLE VI

OCCUPATIONAL INTELLIGENCE LEVELS, BASED ON ARMY PSYCHOLOGICAL TESTS  
OF 36,500 MEN. ALPHA SCALE  
(from Proctor [47])

Occupation	Median Score	Range of Middle 50 Per Cent.
Laborers (unskilled) . . . . .	35	21 to 63
Semi-skilled labor		
Cobblers . . . . .	39	23 to 67
Teamsters . . . . .	41	23 to 68
Farm workers . . . . .	42	24 to 70
Barbers . . . . .	43	22 to 70
Horse-shoers . . . . .	44	25 to 70
Skilled labor		
R.R. shop-mechanics . . . . .	45	26 to 83
Bricklayers . . . . .	48	23 to 81
Cooks . . . . .	49	28 to 79
Bakers . . . . .	53	35 to 83
Painters . . . . .	53	31 to 79
Blacksmiths . . . . .	54	29 to 83
Bridge carpenters . . . . .	55	27 to 84
General carpenters . . . . .	57	33 to 85
Butchers . . . . .	58	33 to 85
Locomotive engineers . . . . .	59	33 to 82
Machinists . . . . .	61	33 to 86
R.R. conductors . . . . .	62	40 to 84
Plumbers . . . . .	62	38 to 87
Tool-makers . . . . .	63	41 to 88
Auto repairmen . . . . .	63	41 to 89
Chauffeurs . . . . .	63	38 to 90
Tool-room-experts . . . . .	64	43 to 88
Policemen—detectives . . . . .	64	41 to 89
Auto-assemblers . . . . .	65	44 to 97
Ship carpenters . . . . .	66	49 to 95
Business and clerical		
Telephone operators . . . . .	70	58 to 99
Concrete construction foremen . . . . .	75	48 to 116
Photographers . . . . .	77	52 to 104
General electricians . . . . .	82	58 to 110
Telegraphers . . . . .	84	59 to 107
R R. clerks . . . . .	92	66 to 116
General clerks . . . . .	96	74 to 123
Mechanical engineers . . . . .	98	63 to 133
Bookkeepers . . . . .	99	78 to 126
Dental officers . . . . .	106	84 to 130
Mechanical draughtsmen . . . . .	112	79 to 134
Stenographers . . . . .	115	93 to 142
Accountants . . . . .	117	101 to 145
Professional		
Civil engineers . . . . .	125	98 to 147
Medical officers . . . . .	130	101 to 165
Army chaplains . . . . .	150	109 to 173
Engineer officers . . . . .	157	134 to 184

which 50 per cent. of the forty items can be solved correctly by only 20 per cent. of college graduates. Each level has a numerical difficulty value, which expresses its distance from the "absolute" zero



of intellect. Measurement throughout the scale is in terms of a constant unit, the S.D. of a ninth-grade group. Not only can scores be added and subtracted, therefore, but a score of forty-two represents twice as much "intellect" (in terms of what the scale measures) as a score of twenty-one.

The zero point on the CAVD was located by a consensus of judgments, with respect to designated tasks, made by psychologists expert in animal and child psychology. Each judge ranked fifty-six tasks in order of merit in accordance with the degree of intelligence required (in his opinion) to perform it. The tasks selected as requiring a minimum of intellect located the absolute zero of intelligence. Illustrations of tasks presumably requiring zero or almost zero intelligence are, "Will not try to pull off his own fingers or toes"; "Having an object of bitter, nasty taste in his mouth, will spit it out more often than hold it there." From this zero the difficulty value of level A (first level) was measured in terms of the P.E. of "expert opinion." This P.E. was then converted into the common scale unit, *viz.*, the S.D. of a ninth-grade group. It is interesting to note that the distance of level A (lowest level actually used) from zero is twenty-three, while the distance of level Q, the highest level, from zero is only forty-three. Thus in terms of the CAVD scale it is as great a feat to progress from "zero" intelligence to the intelligence of the average three-year-old, as it is to move from this intelligence level to that of a superior college graduate. The growth curve of the CAVD is logarithmic (see Figure 4) and probably reaches its level close to twenty years. The point has been stressed that this is a "true" age-progress curve, since growth is measured from zero in terms of a fixed unit (p. 19).

The CAVD examination yields three measures of intellect: *altitude*, *width*, and *area*. The *altitude* or level score is the highest point reached by an individual on the CAVD scale—his maximum point of progress up the scale from just "zero" intellect. The *width* score—defined as the percentage of successes at one or more levels—gives the breadth or scope of intelligence at a given level or at given levels of difficulty. The *area* score is the total number of tasks done—the total range of intelligence in CAVD throughout the scale. These aspects or phases of mental ability are closely correlated, almost identical in fact, the averages of their intercorrelations when corrected for attenuation being close to unity. Hence, the separation

of altitude from width or width from area of intelligence is a convenient rather than a real distinction. Successive levels of CAVD are closely related, the average correlation between one level and the next being about .85; between a single level and another, one to two steps removed .86.

CAVD measures much the same abstract ability measured by other individual and group tests of general intelligence. In a group of 180 pupils in Grades 7 to 12, CAVD correlated .87 with the Otis Self-Administering Test; .94 with the Terman Group Test of General Intelligence; and .78 with Stanford-Binet. CAVD's superiority to these examinations lies in its greater range and in the constant meaning of its scores throughout this range. CAVD is strictly a "power" test, unlimited time being allowed those taking it.

The CAVD examination includes some non-language material at its lower levels, but throughout the greater part of its range it is a verbal test. The test may be administered either as an individual or as a group examination in the upper levels. At the lowest levels, A and B for example, it must necessarily be given as an individual test. The following levels of CAVD are bound together in separate booklets: A, B, C, D, E for pre-school, primary, and lower elementary grades; F, G, H for elementary and lower high-school grades; I, J, K, L for high-school and lower college levels; M, N, O, P, Q for upper college and graduate levels. These tests can be secured from the Bureau of Publications, Teachers College, Columbia University.

### REPRESENTATIVE GROUP TESTS OF GENERAL INTELLIGENCE

This section contains a selection of representative group tests of general intelligence which have proved especially useful in elementary and secondary grades, and in colleges. This list is suggestive rather than exhaustive. References to tests other than those herein described will be found at the end of the chapter. The choice of a particular test will depend upon what one wishes to use it for, as well as upon such practical considerations as time and money allowance, clerical help available, *etc.*

#### ELEMENTARY GRADES<sup>1</sup>

##### 1. Detroit Alpha Intelligence Test, by Harry J. Baker

*Date:* 1924.

<sup>1</sup>Non-language and performance tests designed for kindergarten and primary grades are described in Chapter II.

*Publisher:* Public School Publishing Company, Bloomington, Illinois.

*Designed for:* Grades 5 to 9.

*Contents:* Eight tests: (1) Information; (2) opposites; (3) classification; (4) block design; (5) generalization; (6) analogies; (7) number relations; (8) disarranged sentences.

*Scores:* Letter ratings, M.A. and I.Q.

*Norms:* Letter ratings and age.

*Time:* About forty minutes.

*Reliability* .91 (retest) for 251 unselected sixth-grade pupils, interval between tests being three months.

2. Haggerty Intelligence Examination, Delta 2, by M. E. Haggerty

*Date:* 1920.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Grades 3 to 9.

*Contents:* Six tests, or exercises, (1) Reading and vocabulary; (2) arithmetic problems; (3) picture completion; (4) same-opposites; (5) practical judgment; (6) information. All of these tests are verbal except (3).

*Scores:* Point scores.

*Norms:* Age and Grade.

*Time:* Thirty minutes.

*Reliability:* Single grade about .60; Grades 3 to 9 about .90.

3. Kuhlmann-Anderson Intelligence Tests, by F. Kuhlmann and R. Anderson

*Date:* 1927.

*Publisher:* Educational Test Bureau, Minneapolis, Minnesota.

*Designed for:* Grades 1 to 12.

*Contents:* Nine series of tests, covering the age range from six to eighteen years. The tests range from simple non-verbal to fairly intricate verbal. Each series of tests is printed in a separate booklet, and is intended for use in a given grade or grades.

*Scores:* M.A. and I.Q.; also Heinis' Mental Growth Units.

*Norms:* Age norms.

*Time:* About one hour; depends upon the age of the group tested.

*Reliability:* Not measured by self-correlation. See: KUHLMANN, F., "The Kuhlmann-Anderson Intelligence Tests Compared with Seven Others,"

*Journal Applied Psychology*, 12:545-594, 1928.

4. Multi-mental Scale, by Wm. McCall, *et al.*

*Date:* 1925.

*Publisher:* Bureau of Publications, Teachers College, Columbia University.

*Designed for:* Grades 2 to 9.

*Contents:* Two forms of 100 sets of words, five words in each set. S is to cross out the one word in each group of five which does not belong with the other four. The relationships to be looked for differ from

sample to sample and are illustrated at the beginning of the test by means of samples.

*Scores:* Point scores are transmuted into T-scores, brightness scores, effort scores, M.A. and I.Q.

*Norms:* Age and Grade.

*Time:* About thirty minutes.

*Reliability:*  $r = .94$ , Grades 3 to 9 inclusive.

5. National Intelligence Tests, Scales A and B, prepared under the auspices of the National Research Council, by M. E. Haggerty, L. M. Terman, E. L. Thorndike, G. M. Whipple and R. M. Yerkes

*Date:* 1920.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Grades 3 to 8.

*Contents:* Scale A: Five tests, (1) Arithmetic reasoning; (2) sentence completion; (3) logical selection; (4) synonym-antonym; (5) symbol-digit.

Scale B: Five tests, (1) Computation; (2) information; (3) vocabulary; (4) analogies; (5) comparison. There are three forms of each scale.

*Scores:* Point scores, M.A., I.Q.

*Norms:* Age and Grade.

*Time:* Twenty-five to thirty minutes each scale.

*Reliability:* Single grade, A or B, about .70; Grades 3 to 8 about .93. Correlation of A and B, Grades 3 to 8, .94.

SECONDARY SCHOOLS AND COLLEGES

1. American Council Psychological Examination, by L. L. and T. G. Thurstone

*Date:* 1924, new forms each year.

*Publisher:* American Council on Education, 26 Jackson Place, Washington, D. C.

*Designed for:* High-school graduates and college freshmen.

*Contents:* Five tests: (1) Completion, (2) artificial language, (3) analogies, (4) arithmetic, (5) opposites.

*Scores:* Point scores and percentiles.

*Norms:* Percentiles for college freshmen on each test and on total; averages for many colleges.

*Time:* Sixty minutes.

*Reliability:*  $r = .95$  for  $N = 300$  freshmen.

2. Army Alpha (Revised) by E. O. Bregman

*Date:* 1925.

*Publisher:* Psychological Corporation, New York City.

*Designed for:* Adults; can be used also in high schools or colleges.

*Contents:* Eight tests: (1) Following directions; (2) arithmetic prob-

lems; (3) best answer; (4) same-opposite; (5) disconnected sentences; (6) number series completion; (7) analogies; (8) information.

*Scores:* Point score and percentiles.

*Norms:* Percentile ratings for the general population.

*Time:* About forty minutes.

*Reliability:* Approximately .95 in large groups.

**3. Otis Group Intelligence Scale (Advanced Examination) Forms A and B, by A. S. Otis**

*Date:* 1918.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Grades 5 to 12.

*Contents:* Ten tests, (1) Following directions; (2) opposites; (3) disarranged sentences; (4) proverb matching; (5) arithmetic; (6) geometric figures; (7) analogies; (8) similarities; (9) narrative completion; (10) memory.

*Scores:* M.A., I.Q., percentiles.

*Norms:* Age and Grade.

*Time:* Sixty minutes.

*Reliability:* .97 for Grades 4 to 8.

**4. Otis Self-Administering Tests of Mental Ability, by A. S. Otis**  
**Intermediate Examination: Forms A, B, C, D**

**Higher Examination: Forms A, B, C, D**

*Date:* 1922.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Intermediate examination: Grades 4 to 9. Higher examination: Grades 9 to 12.

*Contents:* Intermediate and higher: seventy-five items of different kinds, in mixed order, make up each examination.

*Scores:* M.A., I.Q. from chart, percentiles.

*Norms:* Age and Grade.

*Time:* Thirty minutes for whole examination. Minimum directions given at beginning of examination.

*Reliability:* Intermediate:  $r = .95$  for Grades 4 to 9. Higher,  $r = .92$ , Grades 7 to 12.

**5. Terman Group Test of Mental Ability, Forms A and B, by L. M. Terman**

*Date:* 1920.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Grades 7 to 12.

*Contents:* Ten tests, (1) Information; (2) best answer; (3) word meaning; (4) logical selection; (5) arithmetic problems; (6) sentence meaning; (7) analogies; (8) mixed sentences; (9) classification; (10) number series.

*Scores:* Point scores, M.A.; percentile scores by grades.

*Norms:* M.A. for point scores; grade.

*Time:* Thirty to thirty-five minutes.

*Reliability:* .89 for ninth grade,  $N = 132$ .

6. Thorndike Intelligence Examination for High School Graduates, by E. L. Thorndike

*Date:* 1919, three forms each year.

*Publisher:* Bureau of Publications, Teachers College, Columbia University.

*Designed for:* High-school graduates and college freshmen.

*Contents:* There are four parts to this examination: a practice form which contains samples of the various tasks in Parts I and II of the examination; Part I, nine tests of directions, arithmetic computation, arithmetic problems, information, same-opposites, word meaning; Part II, six tests of completion, algebra problems, technical and general information; Part III, eight tests of ability to read and answer questions on different paragraphs.

*Scores:* Point scores.

*Norms:* Average scores for freshmen of different institutions.

*Time:* Total, about three and one-half hours; scoring by key.

*Reliability:*  $r = .85$  for 171 normal school students;  $r = .83$  for 562 Columbia College Freshmen.

### THE REQUIREMENTS OF A GENERAL INTELLIGENCE TEST

There are a number of standards which a satisfactory test of "general" intelligence, whether individual or group, must meet. Some of these deal with the choice and arrangement of material, others with criteria of a statistical sort which the completed test must satisfy. An adequate test of general ability on the abstract or verbal level should sample, in the first place, a large and varied number of mental operations. Secondly, it should deal with the relatively more differentiating operations. Tests of analogies, of arithmetic, or of vocabulary are more diagnostic of abstract mental ability than tests of cancellation or digit span. In the third place, a good intelligence test should sample abilities in which every subject (whether student or otherwise) has had approximately equal opportunity, and in which unequal motivation and special gifts play a relatively small rôle. Tests of opposites, arithmetic, and sentence completion, for instance, tap a more homogeneous background of preparation and interest than tests of musical appreciation, ancient history, or domestic science. The general statistical requirements of an intelligence test may be listed as follows:

## 1. Validity

A valid test is one which measures what it purports to measure. If a test is to be employed for the purpose of estimating aptitude for school work, it would naturally be validated against ability to learn in school. Hence, general intelligence tests, designed for use with students, are validated against school grades, teachers' estimates of the intelligence shown in grasping the material taught in courses, and other more or less objective and common sense gauges of ability (45).

## 2. Reliability

A reliable measuring instrument is one which gives self-consistent results. If a child is weighed ten times on a certain scale and the measures of his weight vary from time to time by very small amounts, a fraction of a pound, for instance, the scales are taken to be fairly reliable. Since all of these measures may be incorrect, however, *i.e.*, invalid, reliability is not the same as validity. In order to validate our scale, we should have to find whether a 100-pound weight, let us say, actually weighs a hundred pounds when placed on the scale. The reliability coefficient of an intelligence test is found by correlating one form of the test against another equivalent form (the retest method); or by correlating one-half of the test, say the odd-numbered items, against the other half, the even-numbered items, and then estimating the self-correlation of the whole test by the Spearman-Brown prophecy formula (22).

## 3. Low Inter-test Correlations

To prevent duplication, the sub-tests of a group test battery should have relatively low correlations with each other. This is an ideal requirement which is met only very approximately by most group

TABLE VII

SHOWING THE CORRELATIONS AMONG THE VARIOUS TESTS IN THE ARMY ALPHA INTELLIGENCE TEST BATTERY. THESE COEFFICIENTS ARE BASED ON THE TEST SCORES FROM ABOUT A THOUSAND RECRUITS

(from *Memoirs*, National Academy of Sciences [43])

No.	Description of Test	1	2	3	4	5	6	7	8
1	Directions . . . . .		.73	.59	.71	.69	.68	.67	.66
2	Arithmetic . . . . .	.73		.75	.79	.76	.77	.71	.74
3	Practical Judgment . . . . .	.59	.75		.81	.75	.61	.67	.78
4	Synonym-Antonym . . . . .	.71	.79	.81		.83	.68	.73	.86
5	Disarranged Sentences . . . . .	.69	.76	.75	.83		.67	.78	.82
6	Number Series . . . . .	.68	.77	.61	.68	.67		.70	.69
7	Analogies . . . . .	.67	.74	.67	.73	.78	.70		.67
8	Information . . . . .	.66	.74	.78	.86	.82	.69	.67	

tests because of the relative homogeneity of most verbal batteries, *i.e.*, tests on the abstract level. This homogeneity may be seen clearly in Table VII, which shows the inter-correlations of the eight tests in Army Alpha. All  $r$ 's are positive, the average for the table being .73.

#### 4. Discriminative Capacity

A good intelligence test should have sufficient range to measure extreme cases and its units should be small enough to separate individuals who do not differ markedly in ability. There should be no zero or perfect score, as neither score gives us any real information about the person tested. In large and unselected groups, scores from a highly discriminative intelligence test will be distributed closely in conformity with the normal probability curve (68).

#### 5. Standardization

An adequate intelligence test should be definitely standardized as to procedure; it should have stable age or grade norms, preferably both; it should be relatively easy to administer and score; its contents, besides sampling a wide range of mental operations, should possess general interest and appeal; and it should not be too long or too costly for extensive use.

### SCORES ON A GROUP TEST OF GENERAL INTELLIGENCE

Ability on a group test of general intelligence may be expressed by various kinds of scores. Several of the most common of these will be described in this section.

#### 1. Point Scores

When the sub-tests in a group examination have been arranged in progressive order of difficulty, the simplest total score is that obtained by adding up the point scores made on the separate tests. Such a total score is obtained from Army Alpha, from the Otis Group Intelligence Examination, the Terman Mental Test, the National Intelligence Test, and many other well-known group tests. Cumulated point scores are open to the criticism that such measures do not ordinarily represent equal units or increments of difficulty. One does not know, for instance, whether an increase of five points in the upper range of scores means the same as an increase of five points in the middle or lower range. It happens, however, that this objection is of no great practical importance except at the extremes



of the range, as the transmutation of point scores into a scale having equal units shows marked discrepancies only at the upper and lower extremes (68). One objection to point scores is that a sub-test which contains a large number of items may receive more weight in the total than a sub-test having fewer items. A further minor objection is that the point score itself is meaningless unless referred to some established norm. This is not true of other types of score. An M.A., an I.Q. or a percentile score, for example, is immediately referable to the child's C.A., the mean I.Q. of 100, or the mean percentile score of 50. But a point score of 145 may be high, low, or medium, depending upon the mean ability of the group. Despite these objections, the point score is a satisfactory measure of ability in most well-constructed standard tests, provided age and grade norms are established. It is meaningful when used with older children and adults (this is not true of the I.Q.), and has ease and simplicity of calculation in its favor.

## 2. Percentiles

In order to express the scores on a group test in percentile equivalents, the distributions of test scores made by ten-year-olds, or college students, or any defined group, must be transmuted into an ability scale which extends from zero to 100 with a mean of 50 (22). The raw or point score corresponding to the tenth percentile, or first decile, is found by counting off 10 per cent. of the distribution from the low end. This raw score, which now becomes 10 in the percentile scale, represents the ability which is exceeded by 90 per cent. of the group. The raw score corresponding to the fiftieth percentile (the median score) represents an ability exceeded by 50 per cent. of the group. Percentile ratings enable one to place an individual immediately with reference to members of his group, e.g., a six-year-old with a percentile score of 30 is exceeded by 70 per cent. of six-year-olds in the given test. Percentiles also have the advantage of enabling one to compare directly the scores made by an individual on tests which are scored in different units (22).

The chief objection to a percentile scale is that it assumes a rectangular distribution, each percentile point being exactly the same percentage of the distribution removed from the one just preceding or just following it. But the distribution of ability as measured by most group tests is normal, rather than rectangular, and hence there

is a greater gap in ability between the eightieth and ninetieth percentiles than between the fiftieth and sixtieth. As was true of point scores, however, discrepancies in percentile units are not serious except at the extremes of the scale. Over the middle of the scale, percentile ratings are quite comparable *inter se*, and have the advantage of ease of interpretation and comparison. In individual comparisons the inequality of units may be unjust, but in the classification of pupils little harm is done.

### 3. M.A. and I.Q. Ratings in Group Tests

In many group tests of general intelligence point scores are transmuted into mental age equivalents. Such mental ages are derived from average scores or norms for the different age groups. If the average score for nine-year-olds on a given test is 84, and the average score for ten-year-olds 108, then a point score of 84 is taken to represent an M.A. of nine, while a score of 108 represents an M.A. of ten. Scores corresponding to fractional year intervals between nine and ten are found by interpolation. From these M.A. equivalents, I.Q.'s are calculated as in individual tests, by dividing the M.A. by the C.A. Of the scoring methods mentioned, group test M.A.'s and I.Q.'s are the least defensible, being often subject to gross errors. There are in the main two reasons for this. In the first place, the growth curves for group intelligence tests are rarely such as to give a theoretically constant I.Q.; and secondly, group tests vary so widely among themselves in length, difficulty, reliability, and content, that I.Q.'s derived from them vary enormously even for the same child.

Let us consider first the question of the growth curve for group tests. The National Intelligence Test may be taken as a good illustration, since its standardization is extremely good and the test is highly reliable. The increase in M.A.<sup>1</sup> from year to year on the N.I.T. for the ages from eight to fifteen is shown by the curves in Figure 7. The middle curve pictures the trend of M.A. against C.A. over the age range covered by the test; the upper and lower curves show the increase over the same age range of the twenty-fifth and seventy-fifth percentiles, respectively. Inspection of these age-progress curves makes it at once apparent that they are almost

<sup>1</sup> M.A. equivalents were taken from Table 6 in the supplement to the 1929 Manual of Directions.

straight lines, and that the distance between the separate curves, *i.e.*, the distribution of M.A.'s, remains approximately constant from age to age. On p. 17 we found that the I.Q. on such a test as the Stanford-Binet remains theoretically constant from year to year because the age-progress curve, in terms of M.A. as unit, is a straight line with *increasing spread at succeeding years*. For a straight-line growth curve, such as we find with most group tests, to return a con-

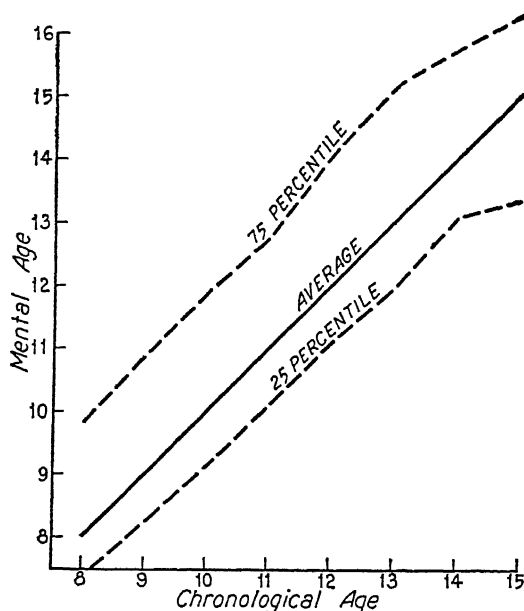


Figure 7.—AGE-PROGRESS CURVES FOR 37,069 CHILDREN UPON THE NATIONAL INTELLIGENCE TEST

stant I.Q., it would be necessary for the lines of growth of dull and bright children (represented by the twenty-fifth and seventy-fifth percentiles in Figure 7) to diverge more and more from the middle ("normal") line as age increases. (See Figure 1, p. 15.) Since this increasing divergence does not occur in the N.I.T., the I.Q.'s from this test cannot remain constant. To illustrate, a child who is one year retarded on N.I.T. at age nine, with an I.Q. of  $8/9$  or 89, would be still one year retarded at age fourteen, but with an I.Q. of  $13/14$ , or 93. In like manner, a bright child two years advanced at age nine, and the same two years advanced at age fourteen, would have

a drop in I.Q. from 122 to 114. Similar results would be obtained with the Otis Group Intelligence Scale (Advanced Examination), as is evident from an examination of Figure 8. It appears, therefore, that the ordinary group test is not adapted to the ratio (I.Q.) technique.

The fluctuation in I.Q. for the same individual from group test to group test has been clearly shown by Gates (24) and Rand (49).

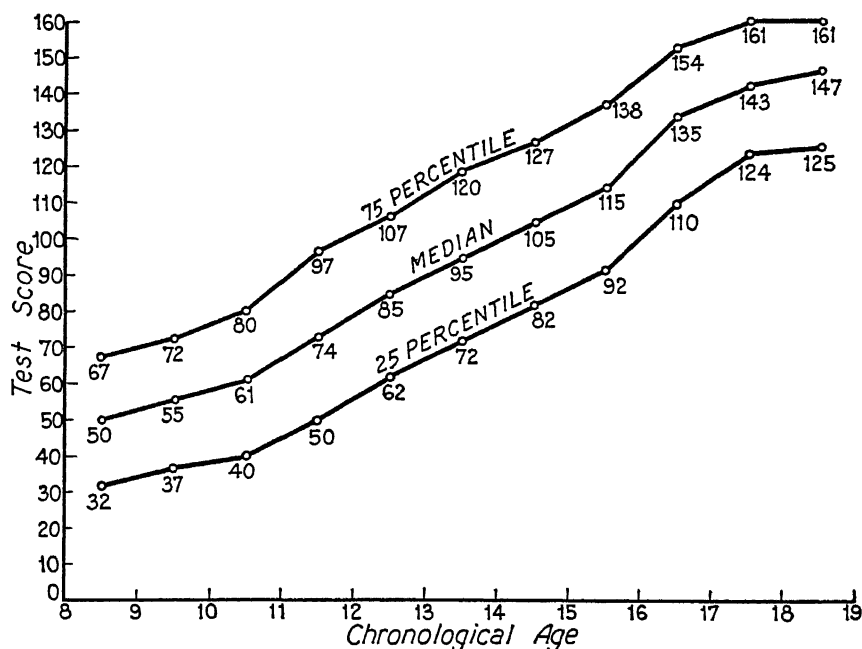


Figure 8.—AGE-PROGRESS CURVES FOR 25,226 CHILDREN UPON THE OTIS GROUP INTELLIGENCE SCALE, ADVANCED EXAMINATION

Gates administered the following group tests to pupils in Grades 1 to 8: Dearborn Examination I or II, Haggerty Delta 1 or 2, Holley's Picture Completion or Sentence Completion Test, the Illinois Intelligence Examination (Grades 3 to 8), Otis Primary or Advanced, N.I.T. Forms A and B (Grades 3 to 8), and the Myers Mental Measure. All of these tests supply age norms from which M.A.'s and I.Q.'s may be calculated. Literally enormous variations in I.Q. were obtained by Gates for the same child when measured on these tests. The range in I.Q.'s received on the different tests by

each of nineteen children in Grade 2, for example, was on the average thirty-five points. One child received an I.Q. on one test seventy-six points higher than that given by another test—the difference between average intelligence and idiocy. In Grade 5 one child earned an I.Q. of 99 on the N.I.T. and 366 on the Myers Mental Measure. Rand's results are less extreme than those of Gates. She presents data to show that the S.D. of the I.Q.'s obtained from a given group test varies so greatly from age to age, and from one group test to another for the same age, as to cause wide variation in the I.Q.'s received by the same child. Suppose a child seven years old has a Stanford-Binet I.Q. of 90, and is 1 S.D. in I.Q. below the mean of his group. Assuming that this child keeps his same position in the other tests *relative* to the group mean, Rand shows that his I.Q. would be 86 on Burt's adaptation of the Binet Test, 81 on Picture Completion Test 2, 72 on the Pressey Primary Test, 70 on the Porteus Mazes, and 80 on the Pintner-Paterson Performance Scale! To quote Rand, ". . . the equivalent to a Stanford-Binet I.Q. of 113 for a high-school freshman would be 113 also on the Terman Group Test, 123 on the Miller Test, 119 on Haggerty, 111 on Otis, *etc.*" Kefauver (34) has suggested a practical remedy for these wide variations in group test I.Q.'s for the same individual. His plan consists of a method for equating I.Q.'s from twelve common group tests. While useful, any such scheme is necessarily a makeshift, for it does not solve the fundamental problem of why I.Q.'s should vary so greatly from test to test.

Wide variations in I.Q. from one group test to another are also contributed to by many extrinsic factors. We have mentioned (p. 47) differences in length, difficulty, content, and reliability among the various group tests. To these we may add wide differences in the size and character of the groups used for standardization purposes; differences in the emphasis upon reading and arithmetic in different elementary schools; and practice with, and habituation to, mental tests.

To summarize briefly the preceding discussion: (1) variations in I.Q. on a single group test are due chiefly to the fact that the age-progress curves of group tests are not adapted to the ratio (I.Q.) technique; in addition, (2) variations in I.Q. from one group test to another arise from differences in length, difficulty, content, and reliability among the various general intelligence tests.

## THE VALUE OF GROUP TESTS OF GENERAL INTELLIGENCE IN SCHOOL WORK

### 1. The Correlation between Group Tests of General Intelligence and Measures of School Achievement

Group tests of general intelligence have been employed in schools for a variety of purposes, all of which, however, center around the selection and classification of children in terms of mental ability. The value of a general intelligence test as a forecaster of school achievement will clearly depend upon the correlation of the test with actual school performance. The correlation of Army Alpha with the amount of schooling reported by men who had been drafted into the army ranged from .50 to about .75 (43). Hundreds of studies have been made of the relationship between general intelligence scores and school achievement in the elementary school, the high school, and in the college. These  $r$ 's run from .30 to .60, depending upon the size of the group, the suitability of the test, and many other factors, the mean correlation being about .45 (45). We may illustrate with a few typical findings. Gates (23) administered several verbal tests of general intelligence in Grades 3 to 8 inclusive. The correlations of the group-test scores with a composite of educational achievement tests ranged from .47 to .65, averaging .54. The number of pupils in each grade was about twenty. Pintner (45) has compiled the correlations reported by fourteen authors between general intelligence test scores and school marks made by high-school pupils. These  $r$ 's, which are based upon varying numbers of students, the largest group being 5,748 high-school seniors, vary from .28 to .60, the mean being .46. Wood (73) has reported correlations of from .45 to .65 between the Thorndike Intelligence Test and freshmen work in Columbia College. Thurstone (70) has reported correlations between the American Council Psychological Examination and college freshmen achievement which range from .40 to .60.

Correlations between various general intelligence tests and college achievement, as reported by different authors, have been compiled by MacPhail (40) and Pintner (45). These correlations cluster around .45. While not especially high, such relationships contrast favorably with the correlations between secondary school grades and

college grades of .26, .33 and .15 reported by Wood (73) on three different occasions.

An interesting prognosis of prospective success in high school based upon scores on Army Alpha has been reported by Cobb (11). Results compiled from many high-school classes in different parts of the country indicated (a) that the minimum Alpha score for entrance to high school at fourteen years is about 65; (b) that a minimum Alpha score of 85 is necessary for a pupil, entering high school at fourteen, to graduate; (c) that the minimum Alpha score for really satisfactory high-school work is around 105. An Alpha score of 65 is roughly equivalent to a Stanford-Binet M.A. of slightly over fourteen years, and to a Stanford-Binet I.Q. of about 90; an Alpha score of 85 to a Stanford-Binet mental age of 15.5 and an I.Q. of 97; an Alpha score of 105 to a Stanford-Binet M.A. of 16.6 and an I.Q. of 104 (78).

The correlation between school grades and intelligence test scores is inevitably limited by many factors. For one thing, school grades are notoriously unreliable. Again, industry, persistence, temperament, character and personality traits, and interests, all enter to a large but undetermined degree into school marks. It is significant, however, that an intelligence test lasting one to two hours will usually give a better forecast of college achievement than grades made in secondary school. This would seem to justify the use of intelligence tests as one criterion, at least, for admission to college. When achievement in the first part of the freshman year is combined with the intelligence test score, future college success is predicted much more efficiently than from the general intelligence test alone. Edgerton (18), for instance, reports an  $r$  of .52 between the Ohio State University Psychological Examination and the first year and a half of college work; this  $r$  becomes .88 when the first quarter's work in college is added to the intelligence test, and .93 when the first and second quarter's work is added to the test. Apparently the odds are distinctly against the satisfactory performance in junior and senior years of a student whose work at the beginning of his college career is definitely poor.

## 2. The Use of Group General Intelligence Tests in the School

We have discussed so far the general value of intelligence tests based upon their correlations with school grades. We shall now

enumerate some more specific uses of the general intelligence test in school work, especially in the elementary and secondary grades.

(a) *Classification of Pupils into Homogeneous Groups:* A good illustration of classification based upon group tests is the Trinidad plan devised by H. M. Corning (14), sometimes known as the "three-track" plan. According to this scheme, which is based upon a plan proposed by Terman (63), the grade location of a child is determined by his mental age. Pupils are then sectioned within the grade into slow, normal, and bright groups on the basis of I.Q. Corning calls this procedure "classifying vertically by M.A. and horizontally by I.Q." The logic underlying this, and somewhat similar classifications, is that mental maturity should determine in what grade a child fits best; while brightness should determine the rate at which a pupil works. Instead of M.A. and I.Q., E.A. and E.Q. (p. 59) as derived from a general achievement test could also be employed. Grouping in terms of mental test score will, in general, prove advantageous both to teachers and to pupils. It is a decided advantage for a teacher to know that the ability in her class is homogeneous and at a given level. And it makes for a better spirit when a pupil is placed where he can work best, so that, for example, the bright pupils are not bored by the constant repetition of familiar material, and the slow pupils discouraged by being continually out of their depth. The only real drawback to a classification of this sort would seem to arise in the case of very bright and very dull children. Unless there are special classes provided for both of these groups, it would seem best to classify the very bright child and the very dull child by chronological age, instead of by M.A., putting the first in the bright group and the second in the slow group. Unless chronological age is considered, both the very bright and the very dull may have social difficulties, either because the other children are all younger, or all older, than the pupil himself.

(b) *Diagnosis of Cause of Failure:* When a child is failing in a given grade or in a school subject it will clear the ground considerably to know whether his group mental test classifies him as bright, normal or slow. The reasons why a bright child fails are usually quite different from the reasons why a dull child fails. Knowledge of mental status and of relative brightness aids the teacher or administrator in making a diagnosis (p. 27).



(c) *Admission to First Grade:* In order to enter first grade it is now agreed that a child should have an M.A. of at least six years (p. 29). If a child's M.A. is six, and his C.A. somewhat below six, admission to first grade should then depend upon satisfactory social and emotional maturity as judged by competent persons.

(d) *Comparison of Grades and Schools:* Intelligence tests are useful in a comparison of grades, of different schools within a system, and of different school systems. Oftentimes the superior performance of a school is a matter of superior student body (e.g., certain private schools) rather than of exceptionally good teaching. Comparison of different schools is put on a fairer basis when the aptitude of the student body for school work is known.

(e) *Educational Guidance:* On p. 29 we have discussed the school progress to be expected from children who possess various degrees of ability as determined by individual general intelligence tests. Several illustrations of how individual general intelligence tests may be employed in the study of pupil difficulties are given on p. 31. These examples show clearly the value of a mental test in the diagnosis of educational difficulty. In making up a pupil's program; in advising a boy to take a commercial rather than a science course, or a girl to take dressmaking rather than geometry; in encouraging a student to go to college or to a trade school, a knowledge of the pupil's abstract level—his general intelligence as given by group or individual test—is exceedingly valuable. When the knowledge of a student's scholastic aptitude is supplemented by a further knowledge of his mechanical aptitude, temperamental traits, social adaptability, and dominant interests, educational guidance becomes no longer a matter of snap judgment but of prediction based upon definite knowledge of potential capacities.

### TESTS OF EDUCATIONAL ACHIEVEMENT

While tests of general intelligence are designed to measure native keenness for academic work, educational tests are designed to gauge achievement in some particular school subject. Educational tests fall into two main divisions: tests of educational achievement, including diagnostic tests, and tests of educational prognosis. The educational achievement test—like the ordinary school examination—measures an educational product: how much arithmetic, or how much history a student knows, or how well he can read. Such tests

measure accomplishment regardless of what the aptitude for achievement may be.

Diagnostic tests, besides measuring achievement, are intended to reveal specific weaknesses in a student's knowledge for the guidance of the teacher. The Compass Diagnostic Tests in Arithmetic (13) and the Gates' Silent Reading Tests (39) furnish good examples in two highly important fields. The first test consists of a series of twenty sub-tests, embracing ninety different skills necessary for successful work in arithmetic. The tests cover the range of arithmetic ability from Grades 2 to 8, and each is sufficiently reliable to permit of individual diagnosis. From this test the teacher can discover in what specific skills a given pupil is lacking or is proficient.

The Gates' Silent Reading Test is a battery of four reading tests covering reading ability in Grades 3 to 8. Each sub-test measures a different and fundamental skill, such as the ability to read quickly for general impression; to read carefully and accurately; to read analytically so as to predict an outcome; and to read for details. The test, as a whole, is designed to furnish a general estimate of reading ability, as well as a diagnosis of specific strengths and weaknesses in reading.

It is obvious, of course, that all achievement tests are in a sense diagnostic. Even when an educational achievement test in French, let us say, is used primarily for survey purposes, *i.e.*, as a measure of level or status, the kind of errors made by a student will indicate to the teacher or examiner whether he is especially weak in vocabulary, grammar, or translation.

The purpose of the educational prognosis test is to estimate beforehand the probable success of a student in a given subject, *e.g.*, Latin or physics. In devising a prognostic test, the effort is made to analyze and measure the essential abilities which make for success in the given subject. As no actual knowledge of the subject matter itself is assumed, such a test measures aptitude rather than accomplishment. An illustration of the prognostic test is the Iowa Chemical Aptitude Examination, one of the series of Iowa Placement Examinations.<sup>1</sup> This test consists of four parts. Part I covers the arithmetic judged to be necessary in chemical calculation. Part II is a para-

<sup>1</sup> Published by the Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

graph-reading test based upon material taken from standard textbooks in chemistry. The idea in this test is to see whether the student can deal with such material accurately and with understanding. Part III is a measure of comprehension in reading fairly difficult prose which deals with topics in chemistry. Part IV attempts to measure the interest already developed by the student in chemistry. Questions cover such general information about chemical subjects as would be gleaned by an interested person from reading popular scientific and other articles on the subject. Other prognostic tests in the Iowa series are the English Aptitude, the Foreign Language Aptitude, Mathematical Aptitude, and Physics Aptitude Tests. These examinations have been found to correlate on the average about .50 with first-semester college grades in the subjects which they cover (57).

If a student has already had some work in a given subject, the ordinary achievement test is definitely prognostic of later performance, on the general theory that future ability in mathematics, or French, say, may be best predicted by past performance in the same subject. In the Iowa Placement Series each aptitude test is paired with a training test in the same subject. These training tests are standardized measures of educational achievement, based largely upon the material required of first-year college students. When the training and aptitude tests are combined, the correlation of the team (for example, Mathematical Aptitude plus Mathematical Training) with first-year grades in the same subject rises to .65, on the average (57). Another series of placement examinations is the Columbia Research Bureau's tests,<sup>1</sup> which cover algebra, geometry, English, French, Spanish, German, history, chemistry, and physics. These tests measure achievement at the high school and college level (usually freshmen, sometimes higher). Extensive norms for all of these tests have been established. Achievement tests are especially valuable, when combined with prognosis tests, in furnishing a basis for recommending special promotion; or the repetition of a course when the grasp of the subject matter is too slight to warrant more advanced work.

It is not the purpose of this chapter to describe in detail the many and various educational achievement tests. Many books are available which deal specifically with educational tests from the standpoint

<sup>1</sup> Published by World Book Company, Yonkers, New York.

of the teacher and administrator. It will be sufficient here to show in what respects the educational achievement test is superior to the ordinary school examination as a measuring device; to indicate how such tests are used in schools; and to supply a list of well-known educational achievement tests as samples of the work in this field.

## CONSTRUCTION OF EDUCATIONAL ACHIEVEMENT TESTS

### 1. Arrangement of Items

In the ordinary school examination little attention is paid to the arrangement of items for difficulty value. The usual plan is to start the examination with fairly easy items, or questions, and end with those which are sometimes more difficult. But this is not always done, and it is not unusual for difficulty to seesaw up and down so that increases in score do not represent equal increases in performance. In constructing a standard achievement test, the difficulty of the separate items is established at the outset, by computing the percentage of a preliminary experimental group which is able to pass each item. In order to eliminate useless and non-discriminating material, each item to be included in the test must show a significantly higher per cent. of passes from one grade to the next. Items passed by 0 per cent. or 100 per cent. of the preliminary group are discarded, as they are of no value in differentiating individuals or groups.

### 2. Objectivity in Scoring

An objective examination is one in which the mark or grade given the student depends to a minimum degree upon the personal opinion of the scorer. In the traditional essay examination a high degree of subjectivity is inevitably introduced because the score on a question depends upon what the individual teacher considers significant and important. Consider the question, "Discuss the causes of the War of 1812," as an example of a common form of essay question. The answers to such a question will contain much material that is true and relevant, and much that is true but irrelevant; much that is false or ambiguous; and much that is clearly "padding." It becomes next to impossible for several people to evaluate the answer in the same way. When answers are recorded by checking an item, circling a number, or underlining a word, the same score is obtained whether it is done by a clerk, or by the teacher of the subject. As

may be readily surmised, the reliability of an examination is greatly increased when the scoring is objective.

### 3. Validity

The routine school examination, in algebra or history, contains questions upon what one teacher regards as the facts worth knowing in one textbook. The content of a standard educational test, on the other hand, is compiled after an analysis of many courses of study, of various textbooks, and of different sets of examination questions. It is a consensus, *i.e.*, the pooled judgment of several competent persons, rather than the judgment of a single individual. A criticism often directed against the validity of the objective examination is that it measures routine information and memory, rather than the ability to organize and evaluate the material taught in the course. The answer to this is that the correlation of objective tests with criteria of school success is usually higher than the correlation of traditional examinations and school success in the same subject. The subject matter of law would certainly seem to require organization and evaluation to a greater degree than routine memory. Yet Wood (74) found, in studying four law courses, an average correlation of .46 between the traditional essay examination and the marks received in the course, as against an average correlation of .76 between objective educational tests and marks received.

### 4. Reliability

Marks assigned to school examinations are notoriously unreliable, when judged by the range of grades given by different teachers to the same persons. Starch and Elliott (55), in a pioneer study of variations in school marks, had 142 teachers of English grade two final examinations in first-year high-school English; 118 teachers of mathematics grade one final examination in geometry; and seventy teachers of history grade one final examination in American history. The variations in grades assigned are little short of astounding. On the two English papers the marks ran from 64 per cent. to 98 per cent., and 50 per cent. to 98 per cent., respectively; on the geometry papers the range was from 28 per cent. to 92 per cent.; and on the history papers from 43 per cent. to 90 per cent. Ruch (51) has summarized in tabular form the reliability coefficients from 285 ordinary school examinations (essay type) reported by different authors. The median reliability coefficient is .59, the range being

from .97 to —.27. By contrast, the reliabilities of standard achievement tests are nearly always above .80 and are often .90 or more. The reliability coefficient of the Stanford Achievement Test (33), for example, was, on the average, .98 in the single age groups from seven to fifteen years (N was approximately 175 at each age.) The reliability of the Iowa High School Content Examination (52) was .95 for 247 high-school seniors; of the Thorndike-McCall Reading Scale (15) .75 for 154 children, Grades 4 to 8; and of the Hotz First Year Algebra Scale (52) .92 for 175 ninth-grade pupils.

A word of caution regarding reliability coefficients is apropos at this point. It must be remembered that a reliability coefficient is meaningful only when considered in connection with the size and variability of the group from which it was obtained. If one simply wishes to estimate the probable future accomplishment of a class from present achievement in a single grade, a reliability of .50 represents a minimum of usefulness. Over three or four grades such a reliability coefficient would increase to .80 or .90. When the range of talent within a group is narrow, the reliability of even a carefully constructed test will be low (33).

### 5. The E.A., E.Q., and A.Q.

The E.A. or educational age is a rating analogous to the M.A. or mental age obtained from an intelligence test, and is derived in the same way. For example, the average score on an educational achievement test made by eight-year-olds represents an E.A. of eight years, the average score made by nine-year-olds, an E.A. of nine years, and so on. When the E.A. is divided by the C.A., the resulting ratio is called the E.Q. or educational quotient. The E.Q. is analogous to the I.Q., and represents the child's educational achievement relative to his life age. Still another ratio, the A.Q., is often calculated. The A.Q., or accomplishment quotient, is the E.Q./I.Q. or the E.A./M.A., and is a measure of the educational achievement of the child relative to his general intelligence or brightness.

Various objections, statistical and otherwise, have been brought against the use of the E.Q. and the A.Q. Since E.A.'s are simply average performances on educational tests, such averages will often differ widely for the same school subject, e.g., arithmetic, because of differences in the length and difficulty of the tests, differences in reliability, and differences in the size and character of the groups

used for standardization. The E.Q., which depends directly upon the E.A., will also vary for the same reasons. Rand (49) has shown that the standard deviation of E.Q.'s is usually smaller than the standard deviation of I.Q.'s calculated from the same groups. This means that even when an E.Q. of 100 corresponds to an I.Q. of 100, the E.Q.'s of children *below* the mean will usually be higher than their I.Q.'s, while the E.Q.'s of children *above* the mean will be lower than their I.Q.'s. To illustrate this effect, suppose that in a given group the S.D. of the E.Q.'s upon a given educational achievement test is 6; and the S.D. of the I.Q.'s upon a general intelligence test is 10. Then if a child is one S.D. *above* the mean in both educational achievement and general intelligence, his E.Q. will be 106 and his I.Q. 110. Instead of the two ratios having the same value, as they should since the child is equally advanced in both tests, the E.Q. is four points less than the I.Q., because of the difference in range. Furthermore, the child's A.Q. is  $106/110$ , or 96 instead of 100 as it should be, since our subject's educational achievement is equal to his general intelligence advancement. On the other hand, if a child is one S.D. *below* the mean, in both the educational and the intelligence tests mentioned above, his E.Q. will be 94 ( $100 - 6$ ) and his I.Q. 90 ( $100 - 10$ ), the A.Q. being 105, instead of 100. Clearly, in such cases, the units of E.Q. and I.Q. are not equal or comparable throughout their range.

The tendency for children of high I.Q. to have somewhat lower A.Q.'s, and for children of low I.Q. to have somewhat higher A.Q.'s, is further shown in the negative correlation usually obtained between I.Q. and A.Q. Popenoe (46) reports a correlation of  $-.46$  between I.Q. and A.Q. for 469 children in Grades 3 to 8. In addition to the error introduced by the ratio method, part of this negative correlation is doubtless owing to the fact that bright younger children are kept working in grades below their mental level while slow older children are pushed ahead. The net result of such a procedure is to lower the A.Q.'s of the brighter children and to raise the A.Q.'s of the duller children. It is worth noting that inequalities in promotion could very well account for this situation, even if the E.Q. and I.Q. scales were equivalent throughout their range.

For an E.Q. to remain constant, the age-progress curve of the educational test from which the E.A. has been obtained should be of the form shown in Figure 1. Makers of educational achievement

tests, however, like the makers of general intelligence tests, have not in general bothered much about the age-progress curves of their tests. To be sure, age and grade norms are usually supplied, and these are sufficient for purposes of comparison and classification. Unless measures of variability at each age are also supplied, however, it is impossible to tell what the form of the age-progress curve is. If the growth curve of an educational test is like that of the N.I.T. shown in Figure 7, it would be impossible for the E.Q. from such a test to remain constant.

The accuracy of an A.Q. depends directly upon the accuracy of the E.A. and M.A. from which it is derived. Since the A.Q. is a ratio, it will reflect all of the errors in both the numerator and denominator. If a child's true I.Q. is 100 and his true E.Q. is 100, the A.Q. will, of course, be 100 also. An error of +10 points in E.Q., however, and of -5 points in I.Q. will give an A.Q. of 116. When the educational and intelligence tests have high reliability; when norms for both of them are obtained from the same population; and when the scales correspond throughout their range, the A.Q. may prove useful as a device for finding whether a child is working up to expectation, or for comparing him with others in his group. But even under these ideal conditions its value is distinctly limited. Thus an A.Q. above 100 is theoretically absurd—for a child obviously cannot do school work at a level above his native capacity. Again, the tacit assumption made in the A.Q. technique, that the one test is measuring educational attainment and the other native ability, is by no means well-founded. Chapman (10) has shown that there is no reliable difference between the measures given by group intelligence tests and educational achievement tests, while Kelley's conclusion that intelligence and achievement tests measure to approximately 90 per cent. the same thing, has been already quoted (33).

## 6. Norms

A decided advantage of the objective educational achievement test over the traditional examination is that age and grade averages are established for all standard educational tests. These norms permit a ready comparison of grade with grade or school with school. They also enable the individual teacher to discover what per cent. of her class is doing work up to the level of the grade, and to compare individual children within the same grade.



DIAGNOSTIC VALUE AND GENERAL USE OF EDUCATIONAL  
ACHIEVEMENT TESTS IN SCHOOL WORK

On p. 52 we described some of the ways in which general intelligence tests may be used in the school situation. These were for purposes of classification, selection, diagnosis of the cause of failure, admission to first grade, comparison of schools and grades, and educational guidance. For most of these purposes the standard educational achievement test will serve as well or better than the general intelligence test (41, 54). Instead of classification by M.A. (p. 53), for instance, we may classify by E.A. Selection and comparison may often be made more profitably in terms of educational achievement, than in terms of general abstract ability. To be sure, there are occasions when the general intelligence test is more useful than the educational test. These are (1) in determining the eligibility of children for admission to the first grade before any formal training has been given; (2) in estimating the probable limit of achievement for a given child early in his school career; (3) in estimating the scholastic aptitude of a child whose education has been "spotty" because of absence due to illness or interrupted training; and (4) as an aid in determining the eligibility of a candidate for admission to college whose background is not of the conventional sort. Intelligence tests are useful, too, when only a short time can be devoted to the estimation of prospective scholastic aptitude.

On the other hand, in order to estimate probable achievement in a single school subject, or the probable general scholastic success of an elementary school child who is entering high school, the educational achievement test is much superior to the general intelligence test. It certainly is more reasonable to estimate future ability in a specific subject by past performance in the same subject, than to estimate it in terms of general ability. Intelligence tests and achievement tests, as has been said above, are to be thought of as supplementary, rather than as measures of different *kinds* of ability. The general intelligence test is the more useful as a measure of general academic expectation; the educational achievement test is by far the better measure of concrete scholastic performance.

## SOME TYPICAL EDUCATIONAL TESTS

The following list of educational tests is a sample selected from various tests in the field. References to many other educational tests suitable for

different grades will be found in the "Tests and Test Materials" at the end of this chapter.

### 1. Achievement Tests

*American Council on Education Achievement Tests* in French, German, Spanish, Civics and Government, Economics, European History, Solid Geometry, Trigonometry. World Book Company, Yonkers, New York. High school and college.

*Buckingham Scale for Problems in Arithmetic*. Public School Publishing Company, Bloomington, Illinois. Grades 3 to 8.

*Columbia Research Bureau Tests* in Algebra, English, American History, French, Chemistry, German, Geometry, Spanish, Physics. World Book Company, Yonkers, New York. High school and college

*Holt Algebra Scale*. Bureau of Publications, Teachers College, Columbia University. First-year algebra.

*Iowa High School Content Examination*. Bureau of Educational Research and Service. University of Iowa, Iowa City, Iowa. High-school seniors and college freshmen.

*Morrison-McCall Spelling Scale*. World Book Company, Yonkers, New York. Grades 2 to 8.

*New Stanford Achievement Test*. World Book Company, Yonkers, New York. Primary examination, Grades 2 to 3; Advanced examination, grades 4 to 9.

*Ruch-Cossmann Biology Test*. World Book Company, Yonkers, New York. High school and college.

*Sones-Harry High School Achievement Test* in Language and Literature, Mathematics, Science, and Social Studies. World Book Company, Yonkers, New York. High school.

*Thorndike-McCall Reading Scale*. Bureau of Publications, Teachers College, Columbia University. Grades 2 to 12.

### 2. Primarily Diagnostic Tests

*Barr Diagnostic Test in American History*. Public School Publishing Company, Bloomington, Illinois. High school.

*Compass Diagnostic Tests in Arithmetic*. Scott, Foresman and Company, Chicago, Illinois. Grades 2 to 8.

*Gates' Silent Reading Tests*. Bureau of Publications, Teachers College, Columbia University. Grades 3 to 8.

*Pressey Diagnostic Tests in English Composition*. Public School Publishing Company, Bloomington, Illinois. Junior and senior high-school grades.

*Sangren-Woody Reading Test*. World Book Company, Yonkers, New York. Grades 4 to 8.

### 3. Prognostic Tests

*Iowa Aptitude Tests* in English, Mathematics, Foreign Languages, Chemistry, and Physics. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa. High school and college freshmen.

*Orleans' Algebra Prognosis Test*. World Book Company, Yonkers, New York. Junior and senior high school.

*Orleans' Geometry Prognosis Test.* World Book Company, Yonkers, New York. Junior and senior high schools.

*Orleans-Solomon Latin Prognosis Test.* World Book Company, Yonkers, New York. Junior and senior high schools.

*Rogers' Test for Diagnosing Mathematical Ability.* Bureau of Publications, Teachers College, Columbia University. Junior and senior high schools.

*Wilkins' Prognostic Tests in Modern Languages.* World Book Company, Yonkers, New York. High school.

### TESTS AND TEST MATERIALS

Catalogues describing a large variety of intelligence and achievement tests, and of test materials, may be secured from the following publishers:

1. World Book Company, Yonkers, New York.
2. Public School Publishing Company, Bloomington, Illinois.
3. Bureau of Publications, Teachers College, Columbia University, New York City.
4. The C. H. Stoelting Company, 424 North Homan Avenue, Chicago, Illinois.
5. The Marietta Apparatus Company, Marietta, Ohio.
6. Southern California School Book Depository, Los Angeles, California.
7. Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.
8. Harlow Publishing Company, Oklahoma City, Oklahoma.
9. Psychological Corporation, 522 Fifth Avenue, New York City.
10. Educational Test Bureau, Inc., University and Fifteenth Avenues, S.E., Minneapolis, Minnesota.
11. C. A. Gregory Company, 345 Calhoun Street, Cincinnati, Ohio.
12. Houghton Mifflin Company, Boston, Massachusetts.
13. Center for Psychological Service, Washington, D. C.

### BIBLIOGRAPHY

1. BINET, A., AND SIMON, T., "L'intelligence des Imbeciles," *L'Année Psychologique*, 1-147, 1909.
2. BINET, A., AND SIMON, T., *The Development of Intelligence in Children.* The Williams and Wilkins Company, Baltimore, Maryland, 1916.
3. BOOK, W. F., *The Intelligence of High School Seniors*, The Macmillan Company, New York, 1922.
4. BREGMAN, E. O., "On Converting Scores on Army Alpha Examinations into Percentiles of the Total Population," *School and Society*, 23: 695-696, 1926.
5. BURKS, B., "The Relative Influence of Nature and Nurture upon Mental Development; A Comparative Study of Foster Parent-Foster Child Resemblance and True Parent-True Child Resemblance," *27th Yearbook, National Society Study Education*, Part 1, 1928.
6. BURT, CYRIL, *Mental and Scholastic Tests*, London, 1921.

7. BURTT, HAROLD, *Principles of Employment Psychology*, Houghton Mifflin Company, Boston, Massachusetts, 1926.
8. CARROLL, H. A., AND HOLLINGWORTH, L. S., "The Systematic Error of Herring-Binet in Rating Gifted Children," *Journal Educational Psychology*, 21:1-11, 1930.
9. CATTELL, P., "Constant Changes in Stanford-Binet I.Q.," *Journal Educational Psychology*, 22:544-550, 1931.
10. CHAPMAN, J. C., "The Unreliability of the Difference between Intelligence and Educational Ratings," *Journal Educational Psychology*, 14: 103-108, 1923.
11. COBB, M. V., "The Limits Set to Educational Achievement by Limited Intelligence," *Journal Educational Psychology*, 13:449-464, 546-555, 1922.
12. COLVIN, S. S., "Principles Underlying the Construction and Use of Intelligence Tests," *21st Yearbook, National Society Study Education*, 33-38, 1922.
13. *Compass Diagnostic Test. Form A.* Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.
14. CORNING, H. M., *After Testing—What?*, Scott, Foresman and Co., Chicago, Illinois, 1926.
15. CURRENT, W. F., AND RUCH, G. M., "Further Studies on the Reliability of Reading Tests," *Journal Educational Psychology*, 17:476-481, 1926.
16. DEARBORN, W. F., *Intelligence Tests*, Houghton Mifflin Company, Boston, Massachusetts, 1928.
17. DICKSON, V., *Mental Tests and the Classroom Teacher*, World Book Company, Yonkers, New York, 1923.
18. EDGERTON, H., *Academic Prognosis in the University*, Maple Press Company, York, Pennsylvania, 1930.
19. FERNALD, M. R., HAYES, M. H., AND DAWLEY, A., *A Study of Women Delinquents in New York State*, The Century Company, New York, 1920.
20. FOX, EDNA, "An Investigation of the Effect of Glandular Therapy on the Intelligence Quotient," *Mental Hygiene*, 12:90-102, 1928.
21. FREEMAN, F. N., *et al.*, "The Influence of Environment on the Intelligence, School Achievement, and Conduct of Foster Children," *27th Yearbook, National Society Study Education*, Part 1, 1928.
22. GARRETT, H. E., *Statistics in Psychology and Education*, Longmans, Green and Co., New York, 1926.
23. GATES, A. I., "The Correlations of Achievement in School Subjects with Intelligence Tests and Other Variables," *Journal Educational Psychology*, 13:129-139, 223-235, 277-285, 1922.
24. GATES, A. I., "The Unreliability of M.A. and I.Q. Based on Group Tests of General Mental Ability," *Journal Applied Psychology*, 7:93-100, 1923.

25. GESELL, A., *Infancy and Human Growth*, The Macmillan Company, New York, 1928.
26. GODDARD, H. H., *A Revision of the Binet Scale*, Training School Bulletin, 8:56-62, 1911.
27. GOODENOUGH, F., *The Kuhlmann-Binet Tests for Children of Pre-School Age*, University of Minnesota Press, Minneapolis, 1928.
28. GRAVES, K., *The Influence of Specialized Training on Tests of General Intelligence*, Teachers College, Columbia University, Contributions to Education, 143, 1924.
29. HERRING, J. P., *Herring Revision of the Binet-Simon Tests*, Teachers College, Columbia University, 1924.
30. HILDRETH, G., "Stanford-Binet Retests of 441 School Children," *Pedagogical Seminary and Journal Genetic Psychology*, 33:365-386, 1926.
31. HULL, C. L., *Aptitude Testing*, World Book Company, Yonkers, New York, 1928.
32. KEAL, H. M., "Mental Ratings, Scholarship and Health," *School and Society*, 28:277-280, 1928.
33. KELLEY, T. L., *Interpretation of Educational Measurement*, World Book Company, Yonkers, New York, 1927.
34. KEFAUVER, G. N., "Need of Equating I.Q.'s Obtained from Group Tests," *Journal Educational Research*, 19:92-101, 1929.
35. KLINEBERG, O., "An Experimental Study of 'Speed' and Other Factors in 'Racial' Differences," *Archives Psychology*, 15:93, 1928.
36. KUHLMANN, F., *A Handbook of Mental Tests*, Warwick and York, Baltimore, 1922.
37. KUHLMANN, F., "A Revision of the Binet-Simon System for Measuring the Intelligence of Children," *Journal Psycho-Asthenics, Monograph Supplement*, no. 1, 1912.
38. KUHLMANN, F., "Results of Repeated Mental Re-examinations of 639 Feeble-Minded over a Period of 10 Years," *Journal Applied Psychology*, 5:191-224, 1921.
39. *Manual of Directions for Gates' Silent Reading Tests, Grades 3 to 8*. Bureau of Publications, Teachers College, Columbia University, 1927.
40. MACPHAIL, A., *The Intelligence of College Students*, Warwick and York, Baltimore, 1924.
41. MCCALL, WM., *How to Measure in Education*, The Macmillan Company, New York, 1922.
42. MEAD, M., "Group Intelligence Tests and Linguistic Disability among Italian Children," *School and Society*, 25:465-468, 1927.
43. *Memoirs*, The National Academy of Sciences, 15; Part 1, 1921.
44. OTIS, A. S., AND KNOLLEN, H. E., "The Reliability of the Binet Scale and of Pedagogical Scales," *Journal Educational Research*, 4:121-142, 1921.
45. PINTNER, R., *Intelligence Testing, Methods and Results*, Henry Holt and Company, Inc., New York, 1921.
46. POPENOE, HERBERT, "A Report of Certain Significant Deficiencies of the

- Accomplishment Quotient," *Journal Educational Research*, 16:40-47, 1927.
47. PROCTOR, W. M., "The Use of Psychological Tests in the Educational and Vocational Guidance of High School Pupils," *Journal Educational Research Monographs*, 1, 1921.
48. PROCTOR, W. M., "The Use of Intelligence Tests in the Educational Guidance of High School Pupils," *School and Society*, 8:473-478, 502-509, 1918.
49. RAND, GERTRUDE, "A Discussion of the Quotient Method of Specifying Test Results," *Journal Educational Psychology*, 16:599-618, 1925.
50. ROGERS, M. C., "Adenoids and Diseased Tonsils; Their Effect upon General Intelligence," *Archives Psychology*, 7:50, 1922.
51. RUCH, G. M., *The Objective or New-Type Examination*, Scott, Foresman and Company, Chicago, Illinois, 1929.
52. RUCH, G. M., AND STODDARD, G. D., *Tests and Measurements in High School Instruction*, World Book Company, Yonkers, New York, 1927.
53. SANDIFORD, P., *Educational Psychology*, Longmans, Green, Ltd., London, 1928.
54. SMITH, H. L., AND WRIGHT, W. W., *Tests and Measurements*, Silver Burdett and Co., New York, 1928.
55. STARCH, D., *Educational Psychology*, The Macmillan Company, New York, 1927.
56. STERN, WM., *The Psychological Methods of Testing Intelligence*, Warwick and York, Baltimore, 1914.
57. STODDARD, G. D., *Iowa Placement Examinations*, University of Iowa Studies in Education, 3:2, 1925.
58. "Symposium: Intelligence and Its Measurement," *Journal Educational Psychology*, 12:123-147, 195-216, 1921.
59. TEAGARDEN, F. M., *A Study of the Upper Limits of the Development of Intelligence*, Teachers College, Columbia University, Contributions to Education, 156, 1924.
60. TERMAN, L. M., *Measurement of Intelligence*, Houghton Mifflin Company, Boston, 1916.
61. TERMAN, L. M., *The Intelligence of School Children*, Houghton Mifflin Company, Boston, 1919.
62. TERMAN, L. M., "The Vocabulary Test as a Measure of Intelligence," *Journal Educational Psychology*, 9:452-466, 1918.
63. TERMAN, L. M., *et al.*, *Intelligence Tests and School Reorganization*, World Book Company, Yonkers, New York, 1922.
64. TERMAN, L. M., *et al.*, "Stanford-Revision of Binet-Simon Scale," *Educational Psychology Monographs*, 18, 1917.
65. THORNDIKE, E. L., *Adult Learning*, The Macmillan Company, New York, 1928.
66. THORNDIKE, E. L., "Intelligence and Its Uses," *Harper's Magazine*, 140: 227-235, 1920.

67. THORNDIKE, E. L., "On the Improvement in Intelligence Scores from 14 to 18," *Journal Educational Psychology*, 14:513-516, 1923.
68. THORNDIKE, E. L., *The Measurement of Intelligence*, Bureau of Publications, Teachers College, Columbia University, 1927.
69. THURSTONE, L. L., "The Absolute Zero in Intelligence Measurement," *Psychological Review*, 35:175-197, 1928.
70. THURSTONE, L. L., AND THURSTONE, T. G., "The 1930 Psychological Examination," *Educational Record*, April, 1931.
71. WALLIN, J. E. W., *Clinical and Abnormal Psychology*, Houghton Mifflin Company, Boston, 1927.
72. WITTY, P. A., AND TAYLOR, J. F., "Some Results of the Multimental Test," *Journal Educational Psychology*, 20:299-302, 1929.
73. WOOD, B. D., *Measurement in Higher Education*, World Book Co., Yonkers, New York, 1923.
74. WOOD, B. D., "The Measurement of Law School Work," *Columbia Law Review*, 25:316-331, 1925.
75. WOODWORTH, R. S., *Psychology; A Study of Mental Life*, Henry Holt and Company, Inc., New York, 1921.
76. YERKES, R. M., AND FOSTER, J. C., *A Point Scale for Measuring Mental Ability*, Warwick and York, Baltimore, 1923.
77. YERKES, R. M., BRIDGES, J. W., AND HARDWICK, R. S., *A Point Scale for Measuring Mental Ability*, Warwick and York, Baltimore, 1915.
78. YOAKUM, C. S., AND YERKES, R. M., *Army Mental Tests*, Henry Holt and Company, Inc., New York, 1920.

## CHAPTER II

### PERFORMANCE AND NON-LANGUAGE TESTS OF GENERAL MENTAL ABILITY

THE present chapter deals with those groups of performance and non-language tests which have been especially designed to measure general mental ability. In Chapter I we found that the verbal intelligence tests measure in large part what we have called abstract or scholastic ability. These tests, in other words, investigate the ability to read and comprehend, to solve problems expressed in words or figures and to grasp relations represented symbolically. In contrast to verbal tests, non-language tests demand a minimum of written or spoken language. Problems presented by means of pictures, diagrams, charts and the like call for ingenuity in perceiving relations, speed of sensory-motor learning, and skill in using non-verbal symbols to arrive at the solution of a problem. In so far as the analysis of content is concerned, language and non-language tests seem to be assaying much the same aspects of ability. But this task is accomplished through the medium of different material.

Performance tests differ from both language and non-language tests in that they require the subject to *do* something rather than to *say* something or to write or mark his answer. Form boards, for example, blocks, picture puzzles, and simple tests requiring memory and manual movement, present concrete problems the solution of which is oftentimes as much a matter of manual activity as of mental alertness. Such tests seem to be measures, predominantly, of the speed and accuracy of manipulative and perceptual activity, although ingenuity and skill in logical selection and discrimination would also seem to be required.

Performance scales and non-language tests have been constructed primarily for use with two groups of subjects: (1) Pre-school children who have not yet acquired written language; and (2) children and adults for whom tests involving written or spoken language are obviously unfair. In this latter group are the foreign-born who do



not speak English; those who stammer or stutter or have other speech defects; the deaf; and those whose use of language is limited because of deprivation from normal contacts or because of a restricted environment. Performance tests are often used, too, in vocational guidance, particularly in the case of those children who find difficulty in performing ordinary school tasks.

In the following section, we shall first consider performance scales designed for use with pre-school children. Because of their simplicity these tests have little application in other groups. Oftentimes, however, they may be used to advantage with low-grade or feeble-minded adults.

### TESTS FOR PRE-SCHOOL CHILDREN

#### 1. Criteria for Selection

It is necessary to exercise more than ordinary care in the selection of tests for use with young children. Such factors as interest, attention span, length of task, fatigue and positive as well as negative incentives have a special importance here and must be carefully considered. It will be valuable, therefore, to list several important criteria which have been set up for the selection of tests for the pre-school child (35).

(a) *Test Material Must Be of Primary Interest to the Child:* Young children like to manipulate things, to make noises and to work with and pull objects about. For this reason simple tasks involving brightly colored objects, blocks, sticks, balls and so forth are most often used as tests. Such things have an immediate appeal to the child's interest, and serve to call forth maximum attention and effort.

(b) *A Wide Range of Activities Should Be Tested:* No one single test can be a fair measure of a child's ability, for even in very young children experience and interests differ widely. A large number of activities, therefore, should be sampled before any conclusion as to the child's level of mental development is drawn. The wider the range and the more varied the tasks employed, the greater is the chance of obtaining a fair measure of a child's abilities and of securing results which will be comparable at different ages.

(c) *Tests Should Measure Fundamental Abilities:* Tests which measure discrimination of color, of size and of shape; judgment and quickness of apprehension; dexterity of movement and motor

control; and adaptability in solving prepared problems are more valuable as measures of mental development than are those which involve routine information or specifically acquired habits.

(d) *Tests Should Be Easy to Give, and as Far as Possible Require Fairly Simple Material*: Test material which is familiar, and which can be manipulated by the child himself, creates a much more favorable situation than more formidable-looking apparatus which is liable to disturb or upset the timid child. The briefer the directions, the longer the time which the examiner may spend in studying the child's reactions.

(e) *Differentiation*: It is important that the tests of a scale discriminate clearly between the different age levels in the case of young children because of the rapid growth changes during the early years. A good test of early mental development is one which shows a steadily increasing percentage of passes as we go up the age scale.

In addition to those criteria which bear directly upon the selection of tests for very young children, there is a general statistical requirement which a completed scale should meet, namely, that its norms be based upon adequate samples. This requirement is especially hard to carry out at the lower age levels. Children in nursery schools or those brought to fairs for "baby contests" are probably somewhat above the average child; while children in institutions or in orphan schools are usually below the average. Ideally, our samples should include members of both these groups, but neither group exclusively. In judging a scale for pre-school children, the adequacy of the samples from the point of view of selection should always be considered.

#### REPRESENTATIVE INDIVIDUAL SCALES FOR USE WITH THE PRE-SCHOOL CHILD

Individual scales developed for use with young children are necessarily composed of performance tests. Ordinarily, they consist of systematic observations of a child's voluntary behavior, and of his responses to simple tasks set up by the experimenter. The scales described in this section were designed especially for use with the pre-school child. In this respect they differ from both the Stanford-Binet and the Kuhlmann-Binet, which, while they include low-level tests for use with very young children (p. 8), cover a much wider

age range than those here described. The Stanford Revision contains a variety of simple performance tasks in age-levels three to six, while the Kuhlmann Revision includes tests which extend to the three-months level. These low-level tests in the Kuhlmann Scale are essentially the same kind as those found in the pre-school scales. Group tests of non-language activities designed for kindergarten and first-grade children are described later on p. 91.

### 1. The Bühler Babytests

Tests for the first and second years of life have been developed in Vienna by Charlotte Bühler (5) and her assistants. Four lines of activity are sampled by these tests: (a) Body control and co-ordination; (b) mental ability as shown by imitation; (c) voluntary control and manipulation, memory and attentiveness; (d) social development, such as smiling when another smiles, frowning when another frowns, actively seeking contacts, manipulation of objects as shown in the development of play, use of toys, active use of materials, *etc.* The first-year series consists of ten tests for each month from two to eleven; the second-year series, of four groups of tests (ten in each group) placed at each three-month period from one year to two years.

The nature of the Bühler Babytests can best be shown, perhaps, by reproducing the tests at a given age level. We have selected the three-month level for illustration. The letters S, B, M and O tell whether the test is regarded by the authors as predominantly a measure of social, bodily, mental or "object-manipulation" development.

#### THREE-MONTH LEVEL

- S 1. Returning a glance with smiling or cooing.
- B 2. Holding head and shoulders erect in a prone position.
- B 3. Flight movements of whole body in response to tactile stimulation.
- M 4. Seeking a source of a sound with eyes.
- M 5. Following moving objects with eyes.
- M 6. Changed reaction upon repeating the presentation of an auditory stimulus.
- M 7. Reaction to the disappearance of a human face.
- M 8. Reaction to mask placed on familiar face.
- M 9. Imitating facial grimaces.
- O 10. Active touch (or feeling) of a flat object.

Although directions for administering the Babytests are quite de-

tailed, considerable training is necessary before one can use the scale effectively. In addition to the tests at the baby's C.A. level, those at the two levels below and at the two levels above are also given. Scoring was first effected by assigning point credit to the tests passed, but M.A. credits are now employed as in the Binet Scales.

The Bühler Babytests give promise of being valuable measures of physical and mental development. The tests were chosen after a preliminary try-out and period of observation, and the allocation of tests at different age levels is reported to be based upon results from about 400 children. More work should be done with them, however. At present their norms are somewhat inadequate, and their standardization is not entirely satisfactory. The reliability of such tests as these is, of course, very difficult to estimate. However, the author reports (5) that retests on twenty-five babies after periods of from four to twelve months were in good agreement with first records.

## 2. Gesell's Developmental Schedules

The Gesell schedules are not tests as usually understood, but rather estimates of development based upon careful observation by trained observers of the child's total behavior (15, 16). Four fields of activity are represented by the items: Motor, language, adaptive and personal-social behavior. The Gesell schedules cover each month from one to ten; there are tests also at twelve, fifteen, eighteen, twenty-one, twenty-four and thirty months, and at four, five and six years. The nature of the kinds of behavior studied may be shown concretely by citing some of the normative items established for the four kinds of activity at different month levels. The following are samples:

(a) *Motor Development*: Makes crawling movements when laid prone on a flat surface (one month); holds head steady when carried or when swayed (four months); raises self to sitting position (eight months); walks with help (twelve months); walks attended on the street (twenty-one months); copies vertical or horizontal lines (thirty months).

(b) *Adaptive Behavior*: Retains definite hold of a ring when it is placed in the hand (one month); turns head in pursuing slowly vanishing object (four months); utilizes handle when lifting inverted cup (eight months); secures a cube wrapped in paper (twelve

months); folds paper once on demonstration (twenty-one months); attempts to build bridge from model (thirty months).

(c) *Language Behavior*: Has differential cries for discomfort, pain and hunger (one month); laughs aloud (four months); vocalizes in interjectional manner (eight months); says two "words" (twelve months); repeats things said (twenty-one months); names five pictures (thirty months).

(d) *Personal-social Behavior*: Shows selective regard for the face (one month); plays in a simple manner with a rattle (four months); pats or smiles at a mirror image (eight months); inhibits simple acts upon command (twelve months); asks for things at the table (twenty-one months); gives full name (thirty months).

A child's performance on an item is scored A+, A, B+, B or C. The letter assigned indicates the relative frequency with which the given behavior is found at the child's age, and hence his degree of development. A+ means that the behavior is found in from 1 to 19 per cent. of children at the given level, and thus indicates advanced development; while C means that the behavior occurs in from 85 to 100 per cent. of children at the given level, and hence indicates a considerably slower developmental rate. The Gesell schedules are carefully selected, sample a wide variety of behavior activities, and are well standardized. No provision is made for obtaining a generalized expression in terms of M.A. or I.Q. from these tests. While this renders the schedules less useful, perhaps, than other scales to the practical psychologist, it permits a truer picture of the regularities and irregularities of development. In the hands of an expert, these tests yield the best summary now available of a child's level of physical and mental development.

The Gesell tests may be purchased from the C. H. Stoelting Company, Chicago, Illinois.

### 3. The Merrill-Palmer Scale of Mental Tests

Investigation which led to the Merrill-Palmer Tests was begun in 1922 under the direction of Helen T. Woolley (35, 36). The completed scale contains ninety-three different tests. These have been arranged in order of difficulty and are applicable to children from one and one-half to six years of age. The tests range from such clearly physical and manipulative activities as building a tower of colored blocks, fitting cut-out pieces into a form board, buttoning a

button and throwing a ball, to simple language tests such as answering questions, repeating words, and giving the "agent" answers in the Woodworth-Wells Action-Agent Controlled Association Test (Chap. IV, p. 113).

Norms for the Merrill-Palmer Scale are based upon 631 children, 300 boys and 331 girls. These children were secured from public and private schools, children's agencies and health clinics, and constitute good samples of the age levels covered by the scale. The order of difficulty of the test items was determined by calculating the age at which exactly 50 per cent. of the children solved a given item correctly. The "age at par," as this point was called, was located, for example, at thirty-three months for the Nest of Cubes Test, and at sixty-nine months for the Manikin Test.

The Merrill-Palmer Scale yields a point score which can be translated into an M.A. equivalent. Percentile equivalents and S.D. equivalents to point scores have also been calculated for each age level. To find the S.D. score for a child we first determine the child's M.A. from his point score. At each chronological age level, the distance in mental months of children who are removed  $\pm .5\sigma$ ,  $\pm 1\sigma$ ,  $\pm 1\frac{1}{2}\sigma$ ,  $\pm 2\sigma$  and  $\pm 2\frac{1}{2}\sigma$  from the mean has been calculated. The S.D. position of the child in his age group can be found, therefore, in terms of his M.A. From this S.D. position and the child's chronological age, an I.Q. may be obtained. Use of the I.Q. is not recommended, however, since the S.D. of test scores at successive six-month intervals shows no consistent increase.<sup>1</sup> To illustrate the usual procedure, suppose a child four years old earns a point score of 68 on the test. This score corresponds to an M.A. of fifty-two months, and represents a position in the child's age group  $+ .5\sigma$  above the mean. The I.Q. of a child forty-eight months old and  $+ .5\sigma$  above the mean of his group is approximately 108.

Directions for administering and scoring the Merrill-Palmer Scale as well as norms of achievement are given in *Mental Measurement of Pre-School Children*, by Rachel Stutsman. Like other individual scales, these tests cannot be given by a novice, and considerable training under supervision is necessary before confidence can be placed in an examiner's results. Test materials, scoring blanks, etc., for the Merrill-Palmer Tests can be purchased from the C.H. Stoelting Company, Chicago, Illinois.

<sup>1</sup> See p. 14 for conditions necessary to give a constant I.Q.

### IMPORTANCE OF TESTING YOUNG CHILDREN

The importance of determining the developmental level of a child at an early age is much greater than is usually supposed. Many physical defects in vision and hearing can be corrected if discovered early, as can also many motor defects, such as poor coördination and awkwardness of movement due to a lack of muscular development. If the child is defective mentally it is well to face the fact early and to adjust the training so that it will lie within the child's grasp. Bad social and personal habits in eating, in caring for eliminative functions and in reacting to people, are better understood and more readily handled when the developmental level of the child is known.

Unhappiness, bordering on tragedy, may result if a child of average, or even somewhat above average, ability, is treated by his parents as though he were a genius, or on the other hand is scolded as being exceedingly stupid. When a child is average or below average mentally, obviously he should not be pushed beyond his depth by ambitious parents. On the other hand, if a child is extremely bright, skillful guidance is necessary if he is to realize his full possibilities. Knowledge of a child's mental, motor and social development enables a parent to adapt his training intelligently, and the better to plan for the future.

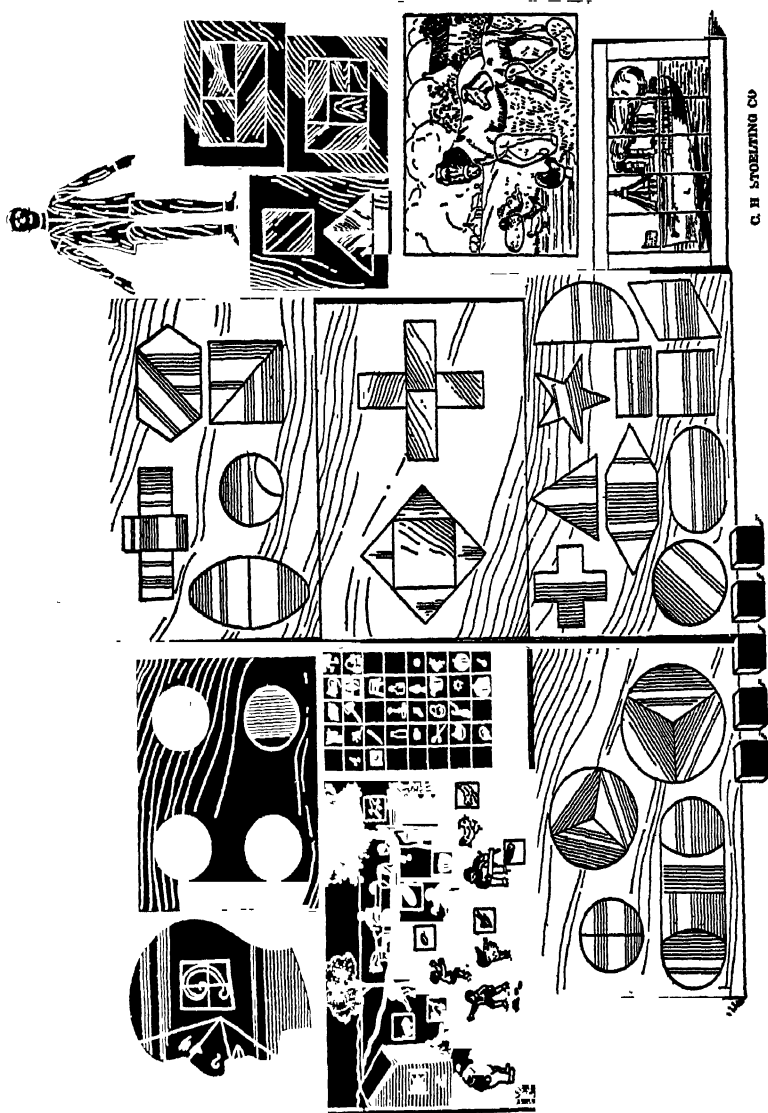
Social agencies are hesitant to offer for adoption young children known to be mentally deficient. Much waste of time, energy and money can be avoided if foster parents know the educational and social possibilities of an adopted child at an early age.

### INDIVIDUAL PERFORMANCE SCALES DESIGNED TO MEASURE GENERAL ABILITY

Most individual performance scales span a much greater age range than those just described, and are applicable to children up to adolescence, and even to retarded adults. Some of the best-constructed and most frequently used performance scales will be described in this section. Many single performance tests are described in Chapters II and III.

#### 1. The Pintner-Paterson Scale of Performance Tests

This scale consists of fifteen manipulative and performance tests which depend to a very slight degree upon the subject's ability to



C. H. STODOLING CO.

Figure 9.—PINTNER-PATERSON PERFORMANCE TESTS  
(Reproduced by courtesy of the C. H. Stodoling Co.)



comprehend and use spoken language (26). Facility in written or spoken language depends partly upon innate capacity, and partly upon home and environmental advantages. Not only does the Pintner-Paterson Scale rule out the obvious advantages accruing to speech from the cultured home, but, more important, it is extremely useful for estimating the abilities of the foreign-born, of the deaf, and of speech defectives. It is useful, too, as a supplement to the Stanford-Binet, when children or adults of the so-called "verbalist" type are encountered, *i.e.*, those in whom a superficial language facility gives a specious impression of a greater degree of abstract intelligence than is actually possessed.

Seven of the fifteen tests in this scale (see Figure 9) are of the form board and three of the picture completion type. In the first group are the Seguin Form Board; the Five Figure Board; the Two Figure Board; the Casuist Form Board; the Triangle Test; the Diagonal Test; and the Healy Puzzle A. These tests are alike in that blocks or cut-out pieces of different sizes and shapes are to be fitted into appropriate depressions in a board. The picture completion tests are the Mare and the Foal; the Ship Test; and the Healy Picture Completion Test, 1. These tests are of the well-known picture puzzle type, and consist of highly colored pictures glued upon a board. Various parts or sections of the picture can be removed in the form of blocks, and these must be fitted into their correct places by the child in order to make the picture complete. The remaining five tests in the Pintner-Paterson series are not easily classified, and may be described individually as follows:

(a) *Manikin Test*: Six wooden pieces, representing the head, the body, the two arms and the two legs of a man, are to be fitted together in correct position. Quality of performance is scored.

(b) *Feature Profile Test*: Eight pieces are to be put together to make a human face. Time taken to complete is scored.

(c) *Substitution Test*: This is the Woodworth-Wells Symbol-Digit Substitution Test described in Chap. IV, p. 94.

(d) *Adaptation Board*: This test is intended to measure the child's ability to follow on a board certain designated movements made by the experimenter.

(e) *Cube Test*: Four cubes placed before the subject are tapped by the experimenter with a fifth cube in a given order. The subject's task is to observe and repeat the experimenter's movements. The score is the number of combinations correctly reproduced.

For general testing purposes a short scale consisting of ten of the fifteen Pintner-Paterson tests is recommended by the authors (25). In this short scale the following tests are omitted: The Triangle Test; the Diagonal Test; the Healy Puzzle A; the Substitution Test and the Adaptation Board. In many respects the short scale is to be preferred to the long scale, since the latter is rather heavily weighted with form board material.

The performance tests of the Pintner-Paterson Scale are scored objectively in terms of time, errors and number of moves—sometimes one and sometimes more than one of these measures being employed. Directions consist of simple demonstrations of what is to be done, followed by the words, "Do this," or "Put this together as quickly as possible."

The age range covered by the Pintner-Paterson series is from four to fifteen years, but no single test is discriminative throughout this range. With normal children the Seguin Form Board will not differentiate beyond age ten; and the Manikin Test beyond age six. Other tests of the series, however, are useful up to twelve or even fourteen years, while the Feature Profile Test has little value before ten years.

In constructing a year scale, the authors have followed the general principle of assigning a test to the age group in which 75 per cent. of those tested pass the test. The limits in time, moves or errors set as the norm for any given age level are the twenty-fifth percentile at that age, and the twenty-fifth percentile of the age next above. If, for example, a seven-year-old child makes a score which falls between the twenty-fifth percentile for seven-year-olds and the twenty-fifth percentile for eight-year-olds—is in the upper 75 per cent. of his own group, or lower 25 per cent. of the age group next above—he is given an M.A. of seven years. Three methods of scoring the tests and of determining a general mental level have been employed: the point scale method; the percentile method; and the median mental age method.

In the point scale method a child is assigned a variable number of points in accordance with his performance on each separate test, and his M.A. is determined from the sum of the points earned on all of the tests. In the percentile method the percentile ratings of a child on the several tests are combined to give a final percentile rating in his age group. In the method of median mental age, a mental age for

each test is determined from established norms, and the median of these several M.A.'s is taken as the final M.A. rating. This last method is the one generally used. Not all of the tests in the scale need be given in order to get a final rating by any one of these methods. This adds considerably to the flexibility of the scale.

While the Pintner-Paterson Scale gives fairly stable and consistent results when scored by a single method, the three methods of scoring do not correspond closely. It is not unusual to secure quite divergent final M.A. ratings for the same child when two or more methods of scoring are used. Considerable scatter, too, will ordinarily appear in the M.A.'s calculated for the same child from the separate tests. Part of this variability is due to the fairly specific functions measured by the test; and part arises from the instability inherent in simple manipulative tests of this sort. Because of this instability, the M.A.'s obtained from separate tests in the series are admittedly rough measures. Chance and luck are decidedly important factors in the simpler form boards; and the puzzle type of test, once solved, is never the same task the second time. For thirteen tests, ten of which were taken from the Pintner-Paterson series, Gaw (14) obtained reliability coefficients of .76 and .54. The first  $r$  was based upon results obtained from a group of ten-year-old boys; the second upon results obtained from a group of ten-year-old girls. When the reliability of the composite is no greater than this, the reliabilities of the separate tests must be quite low. Johnson and Schriefer (21) report that the correlations between M.A.'s obtained from the separate Pintner-Paterson tests, and the median M.A. from the whole scale, range from .13 to .70, with a median at .50. Such a wide range of correlations between separate tests and total score indicates considerable specificity in the abilities measured by the single tests.

The reliability coefficient of the Pintner-Paterson Scale (median mental age method) found by retesting 107 children from three to eleven years old, is reported by Johnson (20) to be .97. This extremely high reliability is in part a function of the wide age range, and would undoubtedly be much lower for single age groups (p. 59).

The Pintner-Paterson Scale is most reliable—and hence most valuable—when used with young children and with retarded or mentally deficient adults. Although few of the tests have discriminative value at the upper age levels, *wide* divergencies from the norms established for the separate tests are probably always significant. But

little confidence can be placed in the small differences which are usually found between groups of the opposite sex, or groups of different racial extraction.

## 2. The Goodenough "Drawing a Man" Scale

The Goodenough "Drawing a Man" Scale (17) measures a child's mental development in terms of his keenness of observation, and his ability to select certain significant items from his environment. The task set by the experimenter is to draw the figure of a man. The human figure was selected because it represents a subject which is familiar and interesting to young children. Directions for the test are simple, the child being told merely to "Make a picture of a man. Make the very best picture that you can." No coaching or suggestions by the experimenter are permitted.

After extensive preliminary trial a scoring scale of fifty-one points was drawn up by means of which the subject's performance is rated. No account is taken of the artistic or lack of artistic qualities in the picture. The child's final score depends upon the presence or absence of such fundamental items as legs, arms, eyes, fingers, nose and mouth in two dimensions; arms and legs attached to the trunk, proportion of parts, *etc.* Upon the basis of results from nearly 4,000 children, age norms corresponding to point scores have been established. A score of ten points, for example, gives the child an M.A. rating of 5.5 years; a score of thirty-four an M.A. rating of 11.5 years. The scale covers a range of ability from 3 to 13 years, but is most effective between the ages of 4 and 10. The author suggests that I.Q.'s may be calculated in the usual way, by dividing M.A. by C.A. This is a dubious procedure, however, since the S.D.'s of the point scores (or of the M.A. norms) do not show the progressive increase from year to year which is necessary to keep the I.Q. constant. Thus, if a child who is  $2\sigma$  above the mean at age four maintains the same degree of superiority at successive ages, his I.Q. will decrease twelve points between four and seven years. At the ages of four to six this child is twenty-one months advanced, his M.A. being 6-3 and his I.Q. 138; while at seven and one-half years he is twenty-four months advanced, his M.A. being 9-6 and his I.Q. 126.

The reliability coefficient of the Goodenough tests is .94 (17). This correlation was determined by retesting 194 first-grade children.

The reliability coefficient for ages five to ten taken separately is .77, on the average. Girls do slightly better than boys on the test, but the sex difference is not marked. The author reports that art instruction as given in the grades has little effect upon the child's score.

### 3. The Kohs Block Design Scale

The Kohs block design scale (22) is a measure of manual activity, as well as of the accuracy and fidelity of perception of likeness and difference. Sixteen colored cubes are employed in this test. Four sides of each cube are painted in white, blue, red and yellow, respectively; the other two sides are divided diagonally, one side being painted half blue and half yellow; the other half white and half red. Before beginning the test, the experimenter examines the subject's recognition of the colors. In administering the test proper a card containing a design made up of white, blue, red and yellow squares is placed before the subject, who is instructed to duplicate this design with the blocks given him. If the child does not understand what he is to do the experimenter completes a trial design, using pantomime. There are seventeen designs in the series, arranged in order of difficulty. Scoring is in terms of time and moves. Each separate and distinct change in the position of a block from its initial position on a table is counted a move.

The Block Design Test was standardized upon a group of 366 children, including both normal and retarded subjects. A child's score is determined by the difficulty values of the designs which he does correctly, supplemented by a record of the time taken to complete, and the number of moves. To illustrate the scoring procedure, if a child does design V correctly in less than thirty-six seconds and in eleven moves, seven is added to his score; if his time is thirty-six seconds to one minute five seconds, one point is subtracted, and if his moves are greater than eleven, another point is subtracted. The final score is converted into a Stanford-Binet mental age equivalent, the probable error of which is sixteen months. This means that in one-half of the cases a child's mental age, as determined by the Block Design Scale, will not differ from his Stanford-Binet mental age by more than one year four months.

The Block Design Scale, like the Goodenough Drawing Scale, measures presumably a rather specialized activity. For this reason it has been most generally employed as supplementary to Stanford-

Binet, or some other language or verbal test. This test may be purchased from the C. H. Stoelting Company, Chicago, Illinois.

#### 4. The Porteus Maze Scale

This series of tests (29) resembles the two scales just described in that it employs only one type of performance, namely, that of tracing the correct pattern through a series of mazes which have been graded for difficulty. There are eleven mazes in the scale. These have been standardized for the ages three, four, five, six, seven, eight, nine, ten, eleven, twelve and fourteen, one maze at each age. The first arrangement of the tests at age levels was drawn up by the author in 1916 as a result of the testing of 1,000 Australian children, five to fourteen years of age. All of these children had been given the Binet tests, and the mazes were validated against Binet M.A. In 1918, as a result of re-standardization upon a group of 1,255 children, several of the tests were revised, and two were discarded and others substituted for them. Although this improved the age assignments of the tests, the scale is still somewhat coarse. Burt (6), who has used the Porteus tests with London school children, suggests that "the mazes as a whole be regarded as forming a single graded test series, roughly increasing in difficulty rather than as marking definite mental ages."

The Porteus mazes are reproduced in Figure 10. The task set in all of the mazes is to trace out with a pencil the correct pathway from entrance to exit. The M.A. is computed by deducting from the highest age at which a test is passed, a year for each test failed at a lower age level, and one-half year for each lower age test passed on a second trial. Instructions for giving the maze tests and specific directions for scoring are given in the monograph: *Guide to Porteus Maze Test*, Publication of the Training School at Vineland, New Jersey, Department of Research, No. 25, March, 1924.

The reliability coefficient of the Porteus mazes based upon re-tests of 110 normal children seven to seventeen years old is reported by Morgenthau (24) to be .95. This very high reliability coefficient must be interpreted with due regard for the wide age range of the children tested.

Porteus has made a strong claim for the maze as a measure of such important social traits as prudence, foresight and planning capacity. The Binet tests he considers to be largely measures of edu-

cational aptitude, and for this reason holds that the mazes are a valuable supplement to the Binet M.A. Porteus' contention is based partly upon an *a priori* analysis of his tests, and partly upon the correlations of his mazes with Binet M.A. and with ratings for industrial, social and educational capacity made upon inmates of the Vineland Institution (29). In a group of twenty-nine feeble-minded males the correlation of Binet M.A. with educational capacity was .64; and with ratings for social and industrial capacity, .50 and .62,

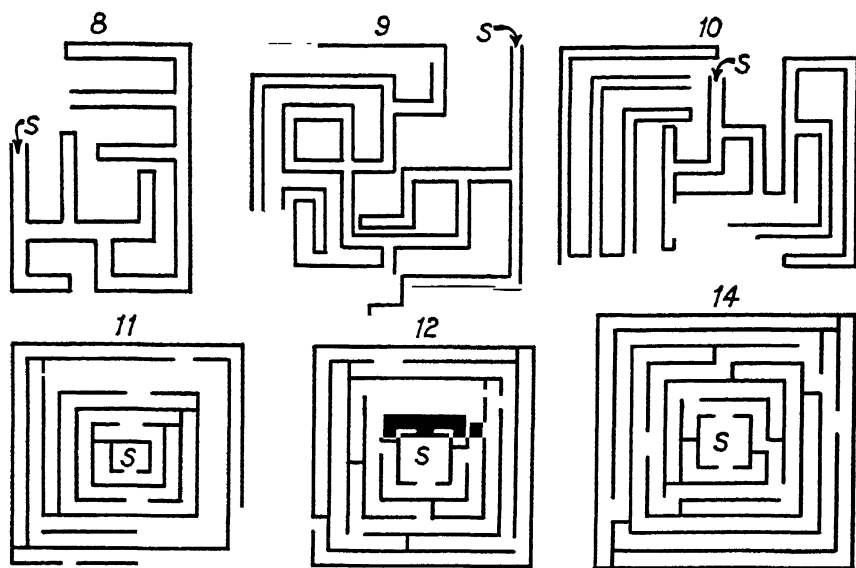


Figure 10.—PORTEUS MAZE TESTS FOR AGES 8, 9, 10, 11, 12, AND 14  
(Reproduced by courtesy of the C. H. Stoelting Co.)

respectively. In the same group the correlation of the Porteus mazes with educational capacity was only .27, while with social and industrial capacity the  $r$ 's were .55 and .67. The correlation between Binet M.A. and Porteus M.A. was .21. In a group of forty-four feeble-minded women the correlations of Binet M.A. with educational capacity was .81; with ratings for industrial and social ability .66 and .59. The correlations of the Porteus mazes with educational, industrial and social capacity in the same group were .59, .75 and .73, respectively. Porteus M.A. and Binet M.A. correlated .60.

From the fact that the mazes correlate somewhat higher than

Binet with industrial capacity and social capacity, and from an examination of many individual records, Porteus concludes that his mazes measure aspects of character and temperament not tested by the Binet. While this may be true, the evidence for it is exceedingly meager. To be sure, Burt considers the Porteus mazes to be useful supplements in "estimating social as distinguished from educational efficiency" (6). But just how he knows that social efficiency is the trait measured by the maze is not made clear. The Porteus mazes may be purchased from the C. H. Stoelting Company, Chicago, Illinois.

### 5. Ferguson Form Boards

The Ferguson Form Boards (12) represent an attempt to construct a graded series of manipulative tests which can be used from the

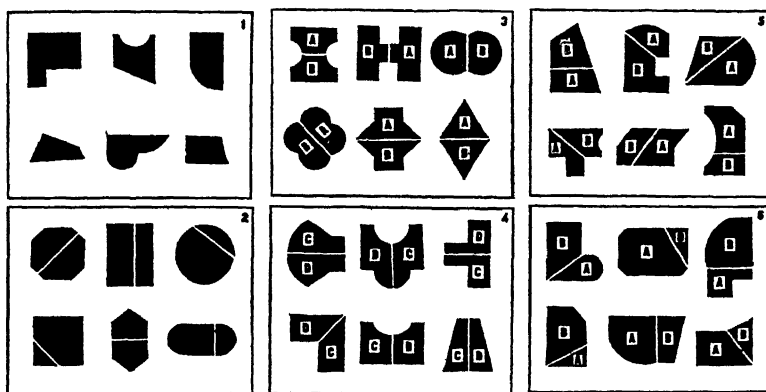


Figure 11.—FERGUSON FORM BOARDS

(Reproduced by courtesy of the C. H. Stoelting Co.)

primary grades to the college level. There are six boards in the series, arranged in approximate order of difficulty by equal steps. The boards are of uniform size and color, each containing six holes cut to receive blocks of irregular sizes and shapes. These boards are represented in Figure 11.

Standardization of these boards was carried out upon 364 subjects ranging from first-grade children to college seniors. The method of administering the test is to present the assembled board to the subject; then empty out the blocks and instruct the subject to replace them as quickly as possible. Scoring is accomplished by assigning from zero to five points to a performance, the number of points assigned depending upon the time taken to assemble the blocks cor-



rectly upon the board. The median scores by grades for the six boards show a steady rise from Grade 1 to the college level. There is also a steady rise in score, when average score for each board is plotted against school grade. The correlation of the Ferguson boards and grade position is .81.

Norms for the Ferguson boards when scored by a slightly different method from the one outlined above may be found in a *Manual of Mental Tests and Testing*, by Healy, Bronner, *et al.* The boards may be purchased from the C. H. Stoelting Company, Chicago, Illinois.

## 6. Miscellaneous Performance Scales

There are several groups of performance tests, in addition to those we have described, which deserve mention.

(a) *Arthur's Point Scale of Performance Tests* (1): This scale comprises two sets of tests. Form 1 contains eight tests taken from the Pintner-Paterson series, plus the Porteus mazes and the Kohs Block Design Test; Form 2 comprises five Pintner-Paterson tests together with the Healy Picture Completion II, the Porteus mazes, and the Kohs Block Design Test. Tables have been drawn up from which performance in terms of time, errors, moves and accuracy may be converted into points. The point total gives the M.A. The Arthur Performance Tests have an actual range of from five years to above fifteen years. Form 2 has been extended by extrapolation to give M.A.'s up to twenty-one years.

(b) *Randall's Island Performance Series* (30): This series of tests is intended for use with children from two to five years of age, and with older defectives. It was designed to supplement the Binet, in order to give a broader picture of ability than that furnished by the verbal tests alone. Tests in the series are classified under nine heads, Manual Planning, Form Perception, Social Orientation, *etc.* The tests themselves have been selected, for the most part, from the Merrill-Palmer Scale, Gesell Schedules, and Pintner-Paterson Performance series. The median of the M.A.'s obtained from the nine activities gives a final M.A. rating.

(c) *The Army Performance Scale* (41): This battery of tests was used with non-English-speaking and illiterate recruits drafted into the army during the World War. There are ten tests in the series. Some of these are non-language tests, such as copying designs, digit-symbol, *etc.*; and some are performance tests, such as the Knox Cube, the Manikin, and the Feature-Profile. Total scores on the scale in

terms of points can be converted into letter ratings, or M.A. equivalents.

(d) *De Sanctis Scale* (8): This scale consists of six simple tests. It was intended by its author to aid in the classification of feeble-minded children, and in the separation of these children from the normal. Although poorly standardized (23), the tests are ingenious and probably possess real value. They measure, mainly, perception and simple manipulative ability.

(e) *Other Performance Tests*: There are several other groups of performance tests which are frequently employed as a scale. One of these is the Worcester Form Board Series (34) which consists of four tests modeled somewhat after the Ferguson boards. Dearborn (7) has also constructed several form boards which are graded in difficulty and used as a battery. Single standard tests are described in Chapters II and III.

#### PERFORMANCE SCALES VS. VERBAL TESTS OF GENERAL INTELLIGENCE

While occasionally used to give a measure of general ability in terms of M.A., I.Q. or point score, performance scales are most frequently employed to supplement the Stanford-Binet or some other verbal or language examination. In this latter use the tacit assumption is made that performance tests measure abilities not reached by the verbal or language intelligence examination. In the present section, we shall examine the evidence for this view. Do performance scales and verbal tests measure different abilities? The answer to this question must be sought in the correlations between the two types of test.

Correlations reported between performance tests and the Binet or its revisions are surprisingly high. The correlation between Pintner-Paterson M.A. and Stanford-Binet M.A. calculated for a group of 488 children, three to thirteen years old, is given by Buford Johnson (20) as .83. A correlation of .74 has been obtained by Goodenough (17) between Stanford-Binet I.Q. and the Drawing a Man Test I.Q. in a group of 334 children, four to ten years old. Kohs reports a correlation of .83 between his Block Design Test and Stanford-Binet M.A. for 366 children from three to nineteen years old. In a group of 190 normal children, age range four to fifteen years, Porteus reports (28) a correlation of .69 between the Goddard Re-

vision of the Binet and the Porteus mazes. In the same study, Porteus obtained a correlation of .77 between Stanford-Binet and the Porteus tests in a group of 263 children of the same age range as the group mentioned above. Correlations of .51, .50 and .56 were obtained by Ferguson (12) between his Form Boards and Army Alpha, teachers' estimates of intelligence and class standing, respectively. Ferguson's subjects were thirty-six sixth-grade children. The correlation between the Merrill-Palmer M.A. and Stanford-Binet M.A. is consistently high (35). On three different trials correlations were .79 for a group of 159 normal children three to six years old; .79 for a group of twenty-nine feeble-minded children four to twelve years old; and .78 for 115 normal children two to six years old. The correlation of the individual tests of the Army performance scale with Stanford-Binet M.A. ranged from .48 to .78 (40), while an abbreviated scale of five tests showed a correlation of .84 with Stanford-Binet. These correlations are based upon groups of adult soldiers varying in size from 134 to 260. Most of these men had failed on Army Alpha or Army Beta.

Taken at face value, these correlations indicate a relatively high community of function between manipulative and performance scales on the one hand, and verbal or language tests on the other. But the evidence, to be entirely convincing, must take account of the wide age ranges within the groups tested. Older children, provided they are normal, will always perform better on mental and physical tests than younger children; not only do they read more rapidly, know more arithmetic, and comprehend more readily, but they are quicker and more skillful with their fingers and hands, taller and stronger. Tests which have been standardized by age levels, *i.e.*, expressly constructed to give progressive increases in score with increase in C.A., must show correlation over a wide age range, even though the abilities measured by the tests are largely independent. Within a single age group, for example, the correlation between general intelligence and height is zero, or negligible. But over a wide age range, scores on a general intelligence examination and measures of height will always correlate substantially, because older children are not only taller than younger children, but they make larger scores. In order to study the essential community of function between two tests, therefore, maturity (age), at least, must be controlled.<sup>1</sup>

<sup>1</sup>It is also desirable to control sex, race and social background.

When maturity is allowed for, the correlation between performance scales and verbal tests drops considerably. Johnson and Schriefer (21) obtained a correlation of .82 between the Pintner-Paterson series and Stanford-Binet M.A. in a group of eighty-six children, three to nine years old. Within the same group, however, the correlation between C.A. and Pintner-Paterson was .78, and the correlation between Stanford-Binet and C.A. .95. If we calculate the correlation between Stanford-Binet M.A. and Pintner-Paterson scale with age variability "partialled out," *i.e.*, held constant, the relationship between the two tests drops from .82 to .40. This correlation of .40 agrees closely with the correlations of .41 and .49 obtained by Gaw (14) between fourteen performance tests, ten of which were taken from the Pintner-Paterson scale, and Binet M.A., in a group of fifty-two boys and in a second group of forty-eight girls, average age of both groups 13.5 years. It is also fairly close to the  $r$  of .51 obtained by Ferguson between his Form Boards and Army Alpha in a sixth-grade group. The age range is narrow in both of these latter investigations.

Kohs has reported the correlation between his Block Design Test and C.A. to be .66 within a group of 291 public-school children, six to seventeen years old (22); and the correlation of the Merrill-Palmer Scale with C.A. was .92 for 631 children one and a half to six and a half years old (35). Porteus does not give the correlation between his mazes and C.A., but since this test is an age-scale, we may safely take its correlation with chronological age to be as high as that of Kohs'. If, now, we take the correlation of Stanford-Binet with C.A. to lie between .85 to .95 (the exact value will depend upon the size and spread in the group), the correlations of such performance scales as Kohs, Porteus and Merrill-Palmer with Stanford-Binet may be taken as falling, roughly, between .45 and .65, for single age groups. It seems probable that the high correlation of .84 between the five tests of the Army performance scale and Stanford-Binet may be attributed to the fact that the men taking the tests were all so low-grade that the "language parts" of the Stanford-Binet did not have a chance to function. Only the lower end of the Stanford-Binet scale was effective. The  $r$ 's of the Goodenough Drawing a Man Test and Stanford-Binet have been calculated separately for each age group from four to ten years. These  $r$ 's range from .56 to .86 with a median at .72. Since maturity does not enter into these  $r$ 's, a cor-

relation of .72 would seem to represent the community of function between the Goodenough test and Stanford-Binet when the variability due to maturity is eliminated.

The discussion in the last paragraphs leads to the tentative conclusion that—when maturity is constant—the overlap of function between the ordinary performance scale and a standard language or verbal test such as Stanford-Binet can be expressed by a correlation of from .40 to .75. Probably .40 is closer to the typical value than .75. To a large degree, therefore, it appears that performance scales are measuring abilities not tapped by the Stanford-Binet, and that their use by clinical psychologists as supplements to verbal tests is justified. This conclusion is strengthened by the low correlations regularly obtained between separate performance tests and verbal tests in groups in which the age-range is narrow. To cite a few instances, in a group of 113 boys, thirteen years and nine months to fourteen years and one month of age (11), the  $r$ 's of the Porteus mazes, Knox Cube Test, Woodworth-Wells Substitution Test, Picture Completion II, Dearborn Form Board, and Cube Construction with a group of verbal tests (opposites, analogies, composition and the like) were .23, .31, .53, .26, .14 and .34—the average being .30. The correlations for the same six tests and the verbal test in a group of 137 girls, of the same age range, were .33, .44, .42, .39, .43 and .44, averaging .41. Worthington (39) reports  $r$ 's from .41 to .79 between single performance tests and Stanford-Binet M.A. in fairly large groups. Turner (38) found correlations of .11, .06 and .13 between Healy Puzzle A and the Otis Group Intelligence Examination, Terman Vocabulary Test, and Trabue Language Completion Scale A. These  $r$ 's were obtained from a group of 108 boys in the eighth grade. The same investigator obtained an  $r$  of .21 between the Trabue Completion Language Scale A and a composite of four form-boards of the Pintner-Paterson series: the Five Figure, Two Figure, Casuist and Triangle. This  $r$  was based upon results from 115 eighth-grade boys. Ross (32) has reported a correlation of .45 between Stanford-Binet M.A. and three performance tests: Mare and Foal, Construction A and Construction B. These correlations indicate very little overlapping of function between single performance tests and language tests.

In concluding this section, let us compare the correlations of performance tests (taken singly or in groups) with verbal tests and the

correlation of the Herring Revision of the Binet scale (p. 11) with the Stanford-Binet. The Herring Revision, like most of the performance scales, was validated against the Stanford-Binet. But it differs radically from the performance tests, being almost entirely abstract and linguistic in character. The  $r$ 's between the Herring and the Stanford-Binet have been computed for each age separately from eight to thirteen, so that maturity is a constant factor. These  $r$ 's range from .97 to .99 with an average at .98 (19). They are based upon from seven to forty-two cases. The average  $r$  of .98, based upon 127 cases, should be compared with the  $r$  of .40 between Pintner-Paterson performance tests and Stanford-Binet when maturity is constant (p. 89). For practical purposes, the Herring is measuring the same "intelligence" as the Stanford-Binet; while the Pintner-Paterson is measuring a concrete and manipulative type of "intelligence" which is, in most respects, different.

#### NON-LANGUAGE GROUP TESTS OF GENERAL ABILITY

Non-language or non-verbal group tests of general ability seem, upon logical grounds, to be more nearly measures of the "abstract" level than performance tests, since they call for relation finding, generalization and the utilization of experience, with a minimum of actual physical movement. We shall describe in this section some representative non-language group tests; and following this, examine the correlations of these tests with verbal group batteries in order to determine to what extent the two varieties of test seem to be measuring the same abilities.

The list of tests here given is not exhaustive. Its purpose is to acquaint the student with the content of non-language tests by reference to some of the better constructed and more often used examinations. References to other tests in this field will be found in the bibliography at the end of this chapter.

1. Army Beta Intelligence Examination. This test was devised by the Division of Psychology, U. S. Army. (See Figure 12.)

*Date:* 1918.

*Publisher:* Beta was originally published by the U. S. government for use with soldiers. It may be purchased now from the C. H. Stoelting Company, Chicago, Illinois.

*Designed for:* Illiterate and non-English-speaking soldiers.

*Contents:* Seven non-verbal tests: (1) Maze drawing; (2) cube analysis; (3) X-O series, completing a series of X's and O's arranged in

different sequences; (4) digit-symbol; (5) number checking; (6) picture completion; (7) geometric construction. These tests were explained by gesture and pantomime, and by means of a large demonstration board. Procedure was acted out before the group.

*Scores:* Points.

*Norms:* Averages for various groups of illiterate and foreign-born adults.

*Time:* Thirty minutes, approximately.

*Reliability:* Not given, probably as high as that of Alpha (.95) in large groups.

2. Dearborn Group Intelligence Tests, Series 1, by W. F. Dearborn

*Date:* 1920.

*Publisher:* J. B. Lippincott Company, Philadelphia, Pa.

*Designed for:* Grades 1 to 3.

*Contents:* Examination A: consists of seventeen items or item groups—all non-verbal—involving directions, drawing and counting, simple information, everyday knowledge and substitution learning. Examination B: consists of five groups of items involving directions, picture sequence, picture recognition, estimates of distance and substitution learning. Items are not organized into sub-test groups.

*Scores:* Points, M.A., I.Q.

*Norms:* Age.

*Time:* Thirty minutes, approximately.

*Reliability:* Not given by author.

3. Detroit First Grade Intelligence Test, by A. M. Engel

*Date:* 1921.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* First grade.

*Contents:* Ten non-verbal tests: (1) Information; (2) similarities; (3) memory; (4) absurdities; (5) comparisons; (6) relationships; (7) symmetries; (8) designs; (9) counting; (10) directions.

*Scores:* M. A.'s and seven letter ratings based on scores; also percentiles.

*Norms:* Age and grade.

*Time:* Twenty to thirty minutes.

*Reliability:* Not given.

4. Haggerty Intelligence Examination, Delta 1, by M. E. Haggerty

*Date:* 1920.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Grades 1 to 3.

*Contents:* Five non-verbal and one verbal test: (1) Following directions (pictures); (2) copying designs; (3) picture completion; (4) picture comparison; (5) symbol-digit; (6) word comparison. Since tests 5 and 6 employ numbers and words, this list is not suitable for children just entering the first grade.

*Scores:* Points.

*Norms:* Age and grade.

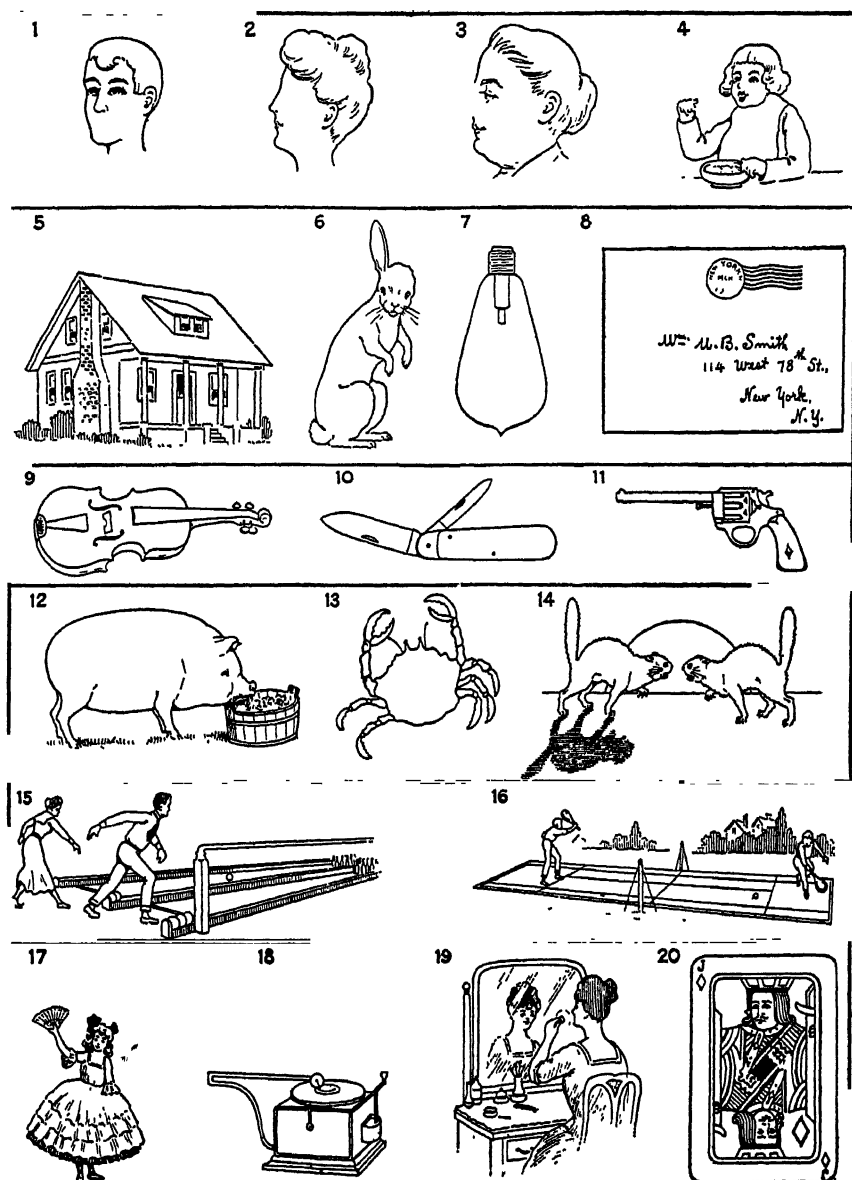


Figure 12.—A SPECIMEN PAGE FROM THE ARMY BETA GROUP EXAMINATION



*Time:* Thirty minutes, approximately.

*Reliability:* .79 (by retest) in group of 100 children in Grades 1A to 2A.

5. International Group Mental Test—Form B, prepared by Stuart C. Dodd and C. C. Brigham under auspices of Committee on Scientific Problems of Human Migration, of the National Research Council (10)

*Date:* 1926.

*Publisher:* Princeton University Press, Princeton, N. J. (sold only for research purposes).

*Designed for:* A wide range of talent from kindergarten to adult, feeble-minded to superior.

*Contents:* Book I: Four non-verbal tests: (1) Cube analysis; (2) picture association; (3) pictorial similarities; (4) similarities in facial expression. Book II: Four non-verbal tests: (5) Maze tracing; (6) pictorial sequence or rhythms; (7) pictorial analogies; (8) picture narratives.

*Scores:* Points.

*Norms:* Means and S.D.'s for various groups: feeble-minded, elementary school children, college groups.

*Time:* Three to four hours for Books I and II and Practice Booklet. About one to one and one-half hours for Book I, or for Book II, when taken separately.

*Reliability:*  $r = .97$  (split-half) for 112 sixth-grade orphans;  $r = .78$  (retest) in same group.

6. Otis Group Intelligence Scale, Primary Examination, Forms A and B, by Arthur S. Otis

*Date:* 1918 (revised editions later).

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Grades 1 to 4, and kindergarten.

*Contents:* Eight non-language tests: (1) Following directions; (2) picture association; (3) picture completion; (4) maze tracing; (5) picture sequence; (6) similarities (pictures); (7) synonym-antonym (simple words); (8) common-sense judgment.

*Scores:* Points, percentiles, M.A. and I.Q.

*Norms:* Age and grade.

*Time:* Twenty-five to thirty minutes.

*Reliability:* Not given in Manual.

7. Pintner-Cunningham Primary Mental Test, by R. Pintner and B. Cunningham. (See Figure 13.)

*Date:* 1923.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Kindergarten and Grades 1 and 2.

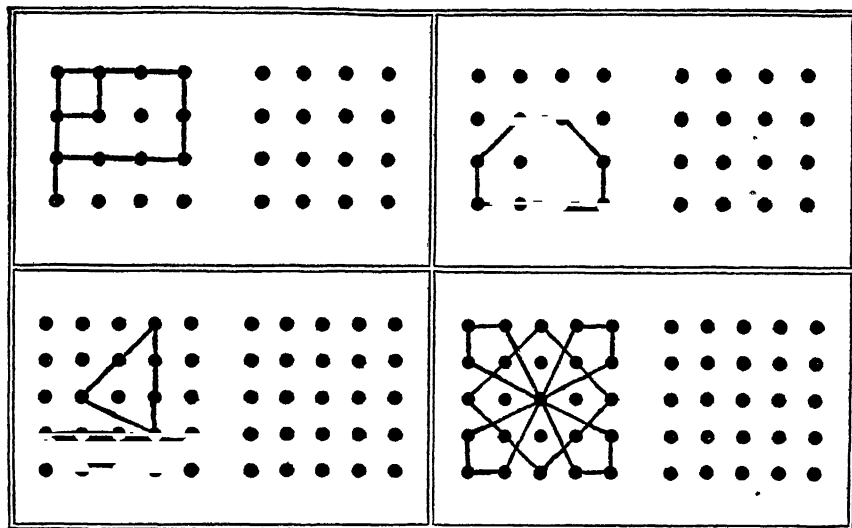
*Contents:* Seven non-verbal tests: (1) Common observation; (2) æsthetic judgment; (3) association of objects; (4) discrimination of size; (5) picture parts; (6) picture completion; (7) dot drawings.

*Scores:* M.A. and I.Q., percentiles.

*Norms:* Age and grade.

*Time:* Thirty to forty minutes.

*Reliability:* About .90 (by retest) for single grade groups (twenty to twenty-five children)



Score.. ..

Figure 13.—PINTNER-CUNNINGHAM PRIMARY MENTAL TEST  
(Specimen Page)

### 8. Pintner Non-Language Mental Test, by R. Pintner

*Date:* 1919.

*Publisher:* College Book Store, Columbus, Ohio.

*Designed for:* Grades 4 to 8.

*Contents:* Six non-verbal tests: (1) Imitation of movement, an adaptation of the Knox Cube Test for group purposes; (2) digit-symbol easy learning; (3) digit-symbol hard learning; (4) picture completion; (5) reproducing reversed figures; (6) picture reconstruction of picture sequence.

*Scores:* M.A., I.Q. and Mental Index in terms of S.D.

*Norms:* Age.

*Time:* Thirty to forty minutes.

*Reliability:* .79 (retest), 201 children, Grades 4 to 6, inclusive.

### 9. Rhode Island Intelligence Test, by Grace E. Bird and C. E. Craig

*Date:* 1923.

*Publisher:* Public School Publishing Company, Bloomington, Ill.

*Designed for:* Kindergarten, ages three to six.

*Contents:* Six non-verbal tests: (1) Marking common objects; (2) omis-

sions; (3) family relations; (4) common activities; (5) social groupings; (6) contrast, similarity and number.

*Scores:* Points.

*Norms:* Age.

*Time:* Fifteen minutes, approximately.

*Reliability:* .92 for 330 children three to six years old.

### VERBAL AND NON-VERBAL GROUP TESTS

The non-verbal or non-language group test was devised primarily with the idea of providing a substitute or equivalent for the verbal test where the latter is inadequate. In deciding whether a child is mature enough mentally to enter the first grade; in selecting and classifying children in the early grades before facility in written language has been acquired; and in vocational problems where illiterates or near-illiterates are to be dealt with, the non-language test fills a real need. Army Beta furnishes a good illustration of the value of the non-verbal test. Alpha, as we know (p. 32), was given to English-speaking enlisted men who could also read and write the language; Beta, on the other hand, was given to foreign-born men who knew little or no English, and to those who used written language so poorly as to be virtually illiterate. The two tests were planned to be at least roughly equivalent, the tasks expressed in language and numbers in Alpha being matched as nearly as possible by tasks expressed in diagrams, charts and pictures in Beta. Test 1, Maze Drawing, in Beta is much like Following Directions in Alpha in that both require the comprehension and carrying through of a series of increasingly difficult tasks. Tests 2 and 3 in Beta, Cube Analysis and X-O Series, probably involve much the same logical and mathematical relation-finding required by Tests 2 (Arithmetic problems) and 6 (Number series completion) in Alpha. Digit-symbol learning, Test 4 in Beta, is not closely analogous to any Alpha test, but Number Checking is matched against Alpha's Synonym-antonym. The Picture Completion Test in Beta represents an attempt to put into concrete form the task involved in the well-known sentence completion test. Picture completion is widely employed in non-language tests as an approximation to such verbal tests as Disarranged Sentences, Analogies and Information. Beta Test No. 7, Geometric Construction, is clearly an attempt to put the form board test on to paper. It does not correspond directly to any test in Alpha.

There are other non-verbal tests which are clearly efforts to parallel language tests. Tests of picture sequence, picture comparison, similarities (in pictures), for instance, are intended to gauge the same mental behavior as the analogies, disarranged sentences, and same-opposite tests. Marking out common objects in a picture, and judgments of size and shape, represent information tests at a low level. Cube analysis, estimating distances, and drawing designs are attempts to measure the ability to use the same sort of quantitative relations as are employed in arithmetic problems, completing number series, and the like.

The correlation of verbal and non-verbal tests in groups of soldiers proved to be surprisingly high. Total scores in Alpha and Beta gave a correlation of .81 in a group of 653 English-speaking, white, enlisted men (40). The  $r$  between Beta and Stanford-Binet M.A. in the same group was .73. One reason for these high  $r$ 's lies, no doubt, in the fact that the men who took Beta were so distinctly low-grade that only the simple parts of Alpha and the Stanford-Binet were attempted. We find, for instance, that 47 per cent. of these men were in the D and E groups on Alpha, and 24 per cent. tested at an M.A. of ten years or below on Stanford-Binet. The distribution of scores on Beta, however, was quite symmetrical. Beta is much easier than Alpha; and in a group of good ability, scores on Beta are all high with a consequent narrowing of the range and poor differentiation.

The intercorrelations of the separate tests in Alpha and Beta are lower than the  $r$ 's between the two tests, but are still fairly high in spite of the shortness of the tests and their consequent low reliability. In a group of 800 enlisted men (37) the average intercorrelation of the eight Alpha sub-tests was .59, with a range from .44 to .75; and in the same group the average intercorrelation of the seven tests in Beta was .44, with a range from .13 to .67. The average intercorrelation of the tests in Alpha and Beta was .41, with a range from .16 to .61. These results show that the separate tests in Beta correlate with the separate tests in Alpha almost as well as they correlate with each other. The two examinations, therefore, have much in common, although they are clearly not identical. Alpha is the more compact and homogeneous of the two tests, the separate Beta tests exhibiting in their intercorrelations more specificity than is shown by the separate Alpha tests. For men of poor ability or of meager educational

attainments it is probable that the two examinations will measure much the same aptitudes.

Correlations almost as high as those between Alpha and Beta or Beta and Stanford-Binet are found between verbal and non-verbal tests in groups of young children. Sangren (33) reports a correlation of .68 between Stanford-Binet and the Pintner-Cunningham Primary Mental Test in a group of 100 first-graders, and a correlation of .75 between the Pintner-Cunningham and the average of seven verbal group tests for the same subjects. Haggerty (18) obtained an  $r$  of .67 in Grades 1 to 3 between his non-verbal group test, Delta 1, and a "verbal" criterion composed of grade location and teachers' ratings for intelligence. The sizes of the groups are not stated. In a group of 144 eight-year-olds, Haggerty reports further a correlation of .64 between Delta 1 and the Haggerty Reading Examination, Sigma 1. Since the age ranges were relatively narrow in these studies, it is probable that the  $r$ 's are not greatly affected by age variability. Rand has reported a correlation of .84 between I.Q.'s on the Dearborn Group Test of Intelligence, Series 1, and Stanford-Binet I.Q. in a group of 211 first-grade children (31). This  $r$  is probably increased somewhat because the correlation is between ratios (I.Q.'s), but the effect is almost certainly not great because of the narrow age range.

Among older children and those who are scholastically more advanced, we should expect the  $r$ 's between non-verbal and verbal tests to be lower than those reported above, because of the failure of most non-verbal tests to discriminate in the upper ability levels. In a group of 108 pupils entering junior high school, Brooks (3) obtained a correlation of .46 between the Pintner Non-language Mental Test and a composite "criterion" of general intelligence, consisting of Stanford-Binet and a battery of nine group tests, such as the Otis, the Terman, *etc.*, and including the Pintner Non-language Test. The correlation of the Pintner Non-language and the Stanford-Binet was .35. Brown (4) has reported the following correlations between the Haggerty Intelligence Examination and the Pintner Non-language Test in groups of boys ten to thirteen years old: age ten,  $r = .55$ ,  $N = 112$ ; age eleven,  $r = .43$ ,  $N = 91$ ; age twelve,  $r = .54$ ,  $N = 120$ ; age thirteen  $r = .59$ ,  $N = 125$ , average  $r = .53$ .

The Princeton International Group Mental Test represents by far the most ambitious attempt to construct a non-verbal test which

will cover a wide range of ability (p. 94). Dodd (9) reports a correlation of .74 between the International and the Stanford Achievement E.A. in a group of 283 orphans in Grades 2 to 6. Since the correlation of the International with C.A. was .71 in the same group, and the correlation of Stanford Achievement E.A. with C.A. almost certainly as high or higher, this correlation is most likely not more than .50 in single age groups. In the Chicago study of the influence of environment on intelligence (13) a correlation of .70 was obtained between Stanford-Binet M.A. and the International, with age variability held constant, in a group of 297 children. Dodd reports the correlation of the International Group Mental Test with school grades to be .72 for 103 twelve-year-olds; and to be .86, with the Thorndike or the Princeton Intelligence Tests, for forty-nine Princeton juniors.

In those groups wherein non-verbal tests ordinarily find their greatest use, namely, young children and low-grade or illiterate adults, their correlations with verbal tests are high enough to make them reasonably good measures of "abstract" ability. The correlations of verbal and non-verbal tests in such groups is fully as high as the correlations of standard verbal group tests *inter se* (p. 44). In groups of better native ability and among those more advanced scholastically, non-verbal tests are, at present, not useful substitutes for verbal examinations. The International Group Mental Test makes it seem probable, however, that non-language group tests can be constructed which will be closely equivalent to the best verbal batteries over a wide range. Such tests should have more value as research instruments, to be used in the comparison of groups having different languages or cultures, than in the prediction of prospective school achievement. For this latter purpose there would seem to be little reason for substituting non-language for language tests.

#### BIBLIOGRAPHY

1. ARTHUR, GRACE, *A Point Scale of Performance Tests*, The Commonwealth Fund, Division of Publications, New York, 1930.
2. BRONNER, AUGUSTA F., HEALY, WILLIAM, LOWE, GLADYS M., AND SHIMBERG, MYRA E., *A Manual of Individual Mental Tests and Testing*, Little, Brown & Co., Boston, 1927.
3. BROOKS, F. D., "The Accuracy of Intelligence Quotients from Pairs of Group Tests in the Junior High School," *Journal Educational Psychology*, 18:173-186, 1927.

4. BROWN, ANDREW W., *The Unevenness of the Abilities of Dull and of Bright Children*, Teachers College, Columbia University, Contributions to Education, 220, 1926.
5. BÜHLER, CHARLOTTE, *The First Year of Life*, The John Day Company, New York, 1930.
6. BURT, CYRIL, *Mental and Scholastic Tests*, London, 1921.
7. DEARBORN, W. F., SHAW, EDWIN A., LINCOLN, EDWARD A., *A Series of Form Board and Performance Tests of Intelligence*, Harvard Monographs in Education, Series I, No. 4, 1923.
8. DESANTIS, SANTE, "Mental Development and the Measurement of the Level of Intelligence," *Journal Educational Psychology*, 2:498-507, 1911.
9. DODD, S. C., *A Statement Concerning Research on an Intelligence Scale for International Use*, 1926, Princeton University Bookstore, Princeton, New Jersey. Offprinted from *International Group Mental Tests* by Dodd, S. C., 1926.
10. DODD, S. C., *International Group Mental Tests*, Princeton, N. J., 1926.
11. EARLE, F. M., MILNER, M., et al., *The Use of Performance Tests of Intelligence in Vocational Guidance*, Great Britain Industrial Fatigue Research Board, No. 53, 1929.
12. FERGUSON, G. O., "A Series of Form Boards," *Journal Experimental Psychology*, 3:47-58, 1920.
13. FREEMAN, F. N., HOLZINGER, K. J., et al., "The Influence of Environment on the Intelligence, School Achievement, and Conduct of Foster Children," *27th Yearbook, National Society for the Study of Education*, Part I, 1928.
14. GAW, FRANCES, "A Study of Performance Tests," *British Journal Psychology*, 15:374-392, 1924-1925.
15. GESELL, ARNOLD, *Infancy and Human Growth*, The Macmillan Company, New York, 1928.
16. GESELL, ARNOLD, *The Mental Growth of the Pre-school Child*, The Macmillan Company, New York, 1925.
17. GOODENOUGH, FLORENCE L., *Measurement of Intelligence by Drawings*, World Book Co., Yonkers, New York, 1926.
18. HAGGERTY, M. E., *Manual of Directions*. Haggerty Intelligence Examination, Delta I, 1929.
19. HERRING, JOHN P., *Herring Revision of the Binet-Simon Tests*, World Book Co., Yonkers, New York, 1924.
20. JOHNSON, BUFORD J., *Mental Growth of Children in Relation to the Rate of Growth in Bodily Development*, E. P. Dutton and Co., Inc., New York, 1925.
21. JOHNSON, BUFORD, AND SCHRIEFER, LOUISE, "A Comparison of Mental Age Scores Obtained by Performance Tests and the Stanford Revision of the Binet-Simon Scale," *Journal Educational Psychology*, 13: 408-417. 1922.
22. KOHS, S. C., *Intelligence Measurement. A Psychological and Statistical*

*Study Based upon the Block Design Tests*, The Macmillan Company, New York, 1923.

23. MARTIN, A. LEILA, *A Contribution to the Standardization of the De-Sanctis Tests*. Training School Bulletin, Publications of the Training School at Vineland, New Jersey, No. 14, 1916.
24. MORGENTHAU, DOROTHY R., "Some Well-known Mental Tests Evaluated and Compared," *Archives Psychology*, 52, 1922.
25. PINTNER, RUDOLF, *Intelligence Testing, Methods and Results*, Henry Holt & Co., Inc., New York, 1931.
26. PINTNER, R., AND PATERSON, D., *A Scale of Performance Tests*, D. Appleton and Co., New York, 1917.
27. PORTEUS, S. D., *Guide to Porteus Maze Test*, Publications of the Training School at Vineland, New Jersey, No. 25, 1924.
28. PORTEUS, S. D., "The Measurement of Intelligence: 653 Children Examined by the Binet and Porteus Tests," *Journal Educational Psychology*, 9:13-31, 1918.
29. PORTEUS, S. D., *Porteus Tests, The Vineland Revision*, Publications of the Training School at Vineland, New Jersey, No. 16, 1919.
30. POULL, LOUISE E., BRISTOL, ADA S., KING, HELEN B., AND PEATMAN, LILLIE B., *The Randall's Island Performance Series*, Columbia University Press, New York, 1931.
31. RAND, GERTRUDE, "The Use of the Correlation Graph with Half-Sigma Class Intervals," *Journal Educational Research*, 9:213-222, 1924.
32. ROSS, ELIZABETH L., "Vocational Tests for Mental Defectives," *Studies in Mental Inefficiency*, 2, 1, 1921.
33. SANGREN, PAUL V., "Comparative Validity of Primary Intelligence Tests," *Journal Applied Psychology*, 13, 394-412, 1929.
34. SHAKOW, D., AND KENT, G. H., "The Worcester Form Board Series," *Pedagogical Seminary*, 32:599-611, 1925.
35. STUTSMAN, R., *Mental Measurement of Pre-School Children*, World Book Co., Yonkers, New York, 1931.
36. STUTSMAN, R., "Performance Tests for Children of Pre-school Age," *Genetic Psychology Monographs*, 1, 1, 1926.
37. THORNDIKE, E. L., "On the Organization of Intellect," *Psychological Review*, 28:141-151, 1921.
38. TURNER, EGBERT M., "Performance Tests as Measures of General Intelligence," *Contributions to Education, New York Society for the Experimental Study of Education*, 2:59-63, 1928.
39. WORTHINGTON, M. R., "A Study of Some Commonly Used Performance Tests," *Journal Applied Psychology*, 10:216-227, 1926.
40. YERKES, ROBERT M., "Psychological Examining in the U. S. Army," *Memoirs of the National Academy of Sciences*, 15, 1921.
41. YOAKUM, CLARENCE S., AND YERKES, ROBERT M., *Army Mental Tests*, Henry Holt and Co., Inc., New York, 1920.



### CHAPTER III

## THE MEASUREMENT OF PERSONALITY AND TEMPERAMENT

THE measurement of traits of personality and of temperament has lagged far behind the measurement of general intelligence and school achievement. There are several reasons for this. In the first place, the terms used to describe non-intellectual traits are broad and often exceedingly general in scope, so that it is hard to secure agreement as to what they mean. Behavior activities included under the head of personality and temperament range all the way from the complex moral or character traits of honesty and loyalty, to the pleasant social habits of courtesy and tactfulness. Secondly, it is difficult to arrive at a satisfactory measure or estimate of a personality trait once it has been defined. It is entirely feasible to set up sample tasks in which the size of an individual's vocabulary or his ability to arrive at a correct conclusion based upon certain premises is determined; or to devise miniature situations in which the speed and accuracy of his responses are measured. But it is extremely difficult to set up an experimental situation in which a person's leadership, or his coöperation, or his honesty can be accurately evaluated.

Perhaps we may most conveniently think of an individual's personality as comprising essentially the sum total of those behavior activities exhibited in social situations. Allport (1) has made the most systematic attempt to analyze personality from this point of view. Under personality he includes *intelligence*, or adaptive behavior; *motility*, by which is meant the general tempo and speed of one's responses; *temperament*, which includes emotional breadth and strength, characteristic moods and attitudes; *self-expression*, concerned largely with social and personal adjustments; and *sociality*, including social participation and the social aspects of character. The present chapter is concerned especially with the measurement of the last three of these activities, intelligence and motor reactions

having been treated in previous chapters. It may be argued with much cogency that *all* mental and physical tests are measures of personality, since results from mental tests depend upon character and temperament as well as upon intelligence and motor skill. This is undeniably true, but it must be remembered that mental tests are administered so as to minimize emotional effects; and that in these tests it is exceedingly difficult to disentangle the emotional from the intellectual components. For these reasons special methods have been devised for measuring personal and social traits, of which there are three in common use. These are (1) The Rating Scale; (2) the Questionnaire; (3) the Objective Test. These techniques will be described in the sections following.

### THE RATING SCALE

The rating scale offers a means of securing quantitative estimates of the degree to which an individual possesses certain abilities or traits. Mary Jones may be rated by her supervisor for ability as a Latin teacher on a seven-point scale, in which 1 indicates "poorest," and 7 "best"; or she may be ranked in order of merit among the Latin teachers of her school. Such ratings as these represent the subjective impressions of the judge, and are not objective measures of performance. Their value will depend upon the number of judges, their accuracy in making estimates, and the degree to which the given trait lends itself to observation and evaluation.

Historically, the rating scale goes back to Francis Galton (21), who in 1883 published a scale for rating the clearness of one's mental imagery. Galton's scale consisted of nine steps or categories, and was used by him to rate his subjects' imagery of their breakfast table. Karl Pearson's (44) scale for rating mental ability was another pioneer feat. This device contained seven steps, intended to run the gamut of intellectual ability. Levels of ability were designated by numbers as follows: (1) Mentally Defective; (2) Slow-Dull; (3) Slow; (4) Slow-Intelligent; (5) Fair Intelligence; (6) Capable; (7) Specially Able. This scale was employed in Pearson's studies of the relation of physical and mental traits, and in investigations upon the inheritance of mental abilities.

The rating scale as a method is related fundamentally to two experimental methods, *i.e.*, order of merit and paired comparisons. In the order of merit method, individuals are put in consecutive

arrangement with reference to some trait, the one possessing the trait in highest degree being ranked first, the one possessing it least being ranked last. In the method of paired comparisons, each individual is compared separately twice over with the other members of his group, once as a standard, and once as a comparison stimulus. The number of preferences for each person determines his position in the final order of merit for the series. These two methods have been shown (3) to be equally good for most purposes. The paired comparisons method has the advantage of permitting the time of the separate judgments to be controlled; but it has the disadvantage of being more tedious and requiring more time than a simple order of merit.

The arrangement of individuals in order of merit becomes exceedingly difficult when their number is large. The very able and the very poor persons are readily identified, but the intermediate positions are often hard to fill. In such cases grouping into larger units is often helpful. Thus, a judge may first classify the objects or persons to be rated into three groups, "Good," "Average" and "Poor," and later rank those within each category separately. If a larger number of groups, say, eight or ten, is used, the group itself may be taken as the unit instead of the individual.

Rating scales may be classified conveniently under four heads: The Man-to-Man Rating Scale; the Graphic Rating Scale; the Numerical Rating Scale and the Descriptive or Adjective Rating Scale. These may be considered in order.

### 1. The Man-to-Man Rating Scale

The man-to-man rating scale was developed by Walter Dill Scott and his associates, and was widely used during the World War under the name Army Rating Scale (45). Officers below the rank of brigadier-general were rated by their immediate superiors for five traits or characteristics: *Physical qualities, intelligence, leadership, personal qualities* and *general value to the service*. The degree to which an officer possessed each of these characteristics was indicated upon a scale containing five steps, each step having a numerical value. We may illustrate with the trait Personal Qualities.

#### IV. Personal Qualities

"Industry, dependability, loyalty, readiness to shoulder responsibility for his own acts, freedom from conceit and selfishness, readiness and ability to coöperate."

Highest .....	15
High .....	12
Middle .....	9
Low .....	6
Lowest .....	3

On the first blank space opposite "highest," the officer making the rating wrote in the name of an officer, well known to him, whom he considered to possess personal qualities (as defined) in the highest degree. In the last space, opposite "lowest," he wrote down the name of an officer whom he considered to possess personal qualities to a minimal extent. The names of five men possessing personal qualities in the highest, high, middle, low and lowest degrees were then written in to fill up the spaces. These five scale men constituted the rating scale for personal qualities, other officers being judged with reference to them. The units of the man-to-man scale were persons known to the rater, and evaluation was made by direct comparison of man and man.

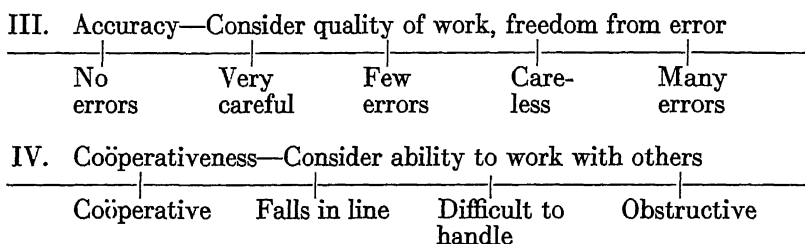
Scale men for the other four traits in the Army Rating Scale were located as described above. After a man had been rated on all five qualities, his ratings were transformed into numerical values and totaled to give a final score. Numerical ratings on each trait ranged from fifteen to three, except general value to the service, in which ratings ranged from forty to eight. Hence, the scale had a range of eighty points, the maximum total being 100 and the minimum twenty.

The man-to-man rating scale has the advantage of concreteness. The probable behavior of an individual, as shown by his possession of certain personal characteristics, is evaluated in terms of another known individual, rather than in absolute terms. But the disadvantages of this type of scale outweigh its advantages. Scott and Clothier (52) contend that the scale proves to be cumbersome and impracticable in business, that the construction of the master scale requires more time and thought than an executive is ready to give, and that it is very difficult to find key men who are satisfactory as scale units. Another objection is that each judge's scale represents his own point of view, and his only. The superior man on one scale or in one trait may be average or low on another scale, or in another trait. If the same individuals could be used as scale men throughout by all of the judges, the scales would then be equivalent. But such agreement is rarely attained and there are usually as many scales as

there are judges. This violates the primary principle of a scale, namely, that it be a constant measuring rod for all those to whom it is applied. When different rating scales are used with the same man, it is as though two examiners measure his height, the one with a yard stick and the other with a meter stick, and compare results directly.

## 2. The Graphic Rating Scale

The graphic rating scale is more flexible than the man-to-man scale and is today probably the most widely used rating technique. In this scheme, traits are usually defined by brief, fairly concrete descriptions. These indicate to the rater what he should consider the term to mean. The scale proper consists of a straight line about four to five inches long, which is taken to represent the range of ability. Beneath the line are written descriptive phrases, intended to define the various points on the scale. We may illustrate with two items from a Graphic Rating Scale for Clerical Workers (5).



The rater places a mark (a check or a cross) somewhere along the line to indicate the degree to which—in his judgment—the person rated possesses the trait in question. The directions make it clear that the cross may be placed *anywhere* along the line, not necessarily at one of the division points. Note that the units on the graphic scale are represented by the divisions on the scale line, and are presumably the same for all judges. Instead of key men who possess certain characteristics, we have specific instances describing varying degrees of the trait.

The graphic rating scale may be scored by assigning numerical values to the different divisions on the rating line. An individual's score then depends upon the distance of the check from the "low" end of the line. If there are five divisions, the lowest may be given the value of one, the highest the value of five, with intermediate values between. The total of credits received on all traits gives the

ratee's final standing in the larger ability of which the separate traits are considered to be a part. The sum of an individual's ratings on the Graphic Rating Scale for Clerical Workers, for example, part of which is shown on p. 108, is intended to represent his value as a clerical worker. Scoring on the Graphic Scale may be facilitated by using a cardboard stencil upon the edge of which divisions are marked off to fit the divisions on the scale line; or by using transparent paper which fits over the whole scale, the lines on it being coincident with the division lines of the scale. Max Freyd (18) has listed the following advantages of the graphic scale method.

"It is simple and easily grasped.

"It is interesting and requires little motivation of the rater.

"It is quickly filled out.

"It frees the rater from direct quantitative terms.

"It enables the rater, nevertheless, to make his discrimination as fine as he cares, although this discrimination is lost if a scoring stencil of only a few points is used.

"It is universal; that is, no master scale is required, as in the Army Rating Scale.

"The fineness of the scoring method may be altered at will, yielding scores of from 1 to 5 or from 1 to 100.

"It allows of comparable ratings without requiring each rater to know all the members of the group."

The American Council on Education Rating Scale (7), shown in Figure 14, is an illustration of a graphic scale.

### 3. The Numerical or Percentage Rating Scale

On a numerical scale, each person is rated on the basis of 100 per cent.; individuals possessing large amounts of a trait are rated 90 per cent., 80 per cent., *etc.*, those possessing small amounts 20 per cent., 10 per cent., *etc.* The chief difficulty with this scale is that it implies much finer differentiation than it yields. The fact that the rating is a number or a per cent. suggests greater definiteness than the scale affords. Moreover, raters seldom agree upon how much 50 per cent. or 20 per cent. of an ability is, so that they employ different units as well as different scales. Because of these drawbacks, the numerical scale has been largely superseded by other rating scales.

### 4. Descriptive or Adjective Scale

The method here is to use common descriptive terms, or broad categories, such as "excellent," "good," "average," "poor," in rating an individual. Thus a teacher's appearance may be rated by

checking "very poor," "poor," "medium," "good," or "excellent"; and a salesman's ability described as "breezy," "cordial," "meets one half way," "reserved," "formal." One variation of the descriptive scale is that in which three categories, "low," "average" and "high" are employed, or two categories "Yes-No." Still another is the use of a series of contrasted adjectives—"Industrious-Lazy," "Accurate-Inaccurate." In these scales the judge describes an individual by checking the more appropriate of two adjectives or by assigning him to a given category.

The use of broad categories is often desirable in rating scales, especially when it is easier to assign an individual to a group or class than to describe him more precisely. If the descriptive phrases used are definite and applicable, such adjective scales may be extremely useful. Like the numerical scale, however, the descriptive scale suffers from the vagueness of its units; nor is it so flexible as the graphic scale. Two judges may both regard a man as excellent, and mean very different things. Again, if a salesman is "reserved" half of the time and "formal" half of the time, this may be indicated on the graphic scale, but on the descriptive scale this person must be assigned forthwith to one category.

The Yes-No descriptive scale and the contrasted adjective forms are open to definite objection. Few persons possess very little or very large amounts of a given ability, since abilities are rarely "all or none," *i.e.*, present or absent. For this reason judges usually find it hard to place a person who does not fall into an extreme grouping, but who occupies an intermediate position.

## THE EVALUATION AND USE OF RATING METHODS

### 1. Construction of a Rating Scale

In preparing a rating scale there are several general principles which may be laid down. First, the selection of traits to be rated is important. As nearly as possible, qualities should be selected which (a) are really valuable; (b) can be exactly defined; (c) are capable of objective evaluation and measurement. The relative value of different traits should be determined by a preliminary analysis of the activities to be covered by the rating scale, and from one's knowledge of what the scale is intended to do. For example, a cultivated voice, tact and sympathy are usually more important to the teacher than to the machinist or manual worker; while leadership, organizing

ability and technical knowledge are more valuable to the executive than to the subordinate.

It is necessary to define traits or qualities on the scale so that raters will not interpret such terms by different standards. If not precisely defined, for instance, one foreman, in rating a worker, may think of "coöperation" as implying a willingness to obey instructions; another foreman as the ability of the man to tie up his work with that of other workers; and still a third as general helpfulness to others. Obviously, the same man rated by these three foremen will show considerable variation in his ratings for "coöperation," simply because of the different meanings attached to the same word. Such general terms as "intelligence," "business knowledge," *etc.*, mean many things to many people, and should be avoided, unless the precise use of the term is indicated by a qualifying description.

Some qualities are more accurately judged and more easily measured than others because they are more open to direct observation. Thus, on a rating scale for manual workers, the descriptive phrase "learns with ease" is more precise than "good" or "satisfactory," or, under the head of "speech" on a rating scale for teachers, the phrase "harsh quality, enunciation indistinct, speech defects" is more illuminating than "poor voice."

Personal characteristics or character traits should be defined, whenever possible, in terms of what the individual rated actually *does* in a given situation, what effect he has on people, *etc.* In constructing the American Council Rating Scale, for instance, Bradshaw (7) sought to objectify the ratings made by requiring that each judge record specific instances ("behaviorgrams") in support of his judgment. Paterson (43) illustrates an objective versus a subjective definition of a trait as follows:

#### APPEARANCE

*Subjective:* Personal attractiveness, cleanliness, neatness, dress.

*Objective:* Consider how favorably he impresses his men by his physique, bearing and manner.

Precise definitions of the *degree* of a trait are important. Thus, "slovenly" is a better term than "very careless in dress," "fastidious" than "very neat," because such terms aptly express certain degrees of the characteristic rated. "Ordinary" or "satisfactory" are bad choices for the middle position of a trait scale, when these terms



imply other and more general qualities than those under consideration. In general, moral or social qualities, or personal habits not ordinarily subject to observation, should not be included on a scale. It is, for example, well nigh impossible for a college president or dean to rate accurately a student's social position, moral habits, or strength of character. Such information is often asked for, however, by teachers' agencies, and, unfortunately perhaps, is often supplied.

In addition to the general principles outlined above, several specific points may be noted which apply in constructing a rating scale, or in evaluating one already constructed. These are as follows:

(a) *The Number of Divisions on the Rating Scale*: In a study of fifty-four teacher rating scales, Boyce (6) found that the number of divisions on these scales varied from two to seven, the mean being four. Symonds (59) has shown that with certain reasonable assumptions as to reliability, the optimum number of rating scale divisions is seven. Many scales, however, contain five divisions, probably because five points correspond to the ordinary marking system, A, B, C, D, E; or to the divisions high, above average, average, below average and poor.

(b) *The Number of Individuals within Each Scale Division*: Unless the group to be rated is quite small, or is known to be exceptional in one or more respects, it is a reasonable assumption that the measurements made upon the members of the group will be distributed in accordance with the normal probability curve. The assumption of normality in trait measurement has many practical advantages. Suppose, for instance, that a class of fifty children is to be rated for "interest in class work" on a five-point scale. Assuming a normal distribution of interest, we should place about 7 per cent. of our group within the first division of the scale; 24 per cent. within the second division; 38 per cent. within the third division (average); 24 per cent. within the fourth division, and 7 per cent. within the fifth and last division. A table giving the percentages of a group which should fall within each scale division, on the assumption of normality, in scales having different numbers of divisions, has been given by Symonds (60).

(c) *The Spread of Ratings on a Given Trait*: If the members of a fairly large group, say forty or fifty, have been rated on a seven-point rating scale, one should ordinarily expect all of the divisions to be used. When only one or two points on the scale are used by a

judge, the average and one other, for example, so little differentiation is secured that the ratings are well nigh worthless. Furthermore, the correlations of such ratings with other measures are of little value. It is clear that no real information is obtained when each of twenty-five teachers is rated "average" or "good" in discipline. Hartshorne and May (26), when using teachers' ratings of their pupils for honesty, threw out all ratings in which only two divisions of a scale (in one case a five-point, in the other a ten-point scale) were used.

(d) *Differences in Standards of Judgment*: Differences in standards of judgment present an especially difficult problem in rating. One judge, for instance, will use the term "excellent" where another uses "good"; a second judge will be unwilling to rate any one at the low end of the scale; still another will employ only the low values, the average being his highest rating. Many authors have noted the general tendency among judges to rate too high (partly "halo effect"), the distribution of ratings being often badly skewed.

Wide variations in ratings arise not only from different standards of judgment, but from conceit, overcaution, prejudice, different degrees of acquaintanceship and failure to understand the meaning of the descriptive terms on the scale. Ambiguities and misunderstandings may be cleared up by careful instructions to the judges. But many difficulties are almost impossible to surmount, and for this reason various methods have been suggested for reducing ratings to a comparable basis. Paterson (43) has suggested the following scheme based upon the assumption that the true distribution of ratings within the group is normal. First, frequency distributions of all of the ratings made by each judge on a single trait are drawn up. Each distribution is then subdivided into five groups: the highest 10 per cent. is designated A; the next 20 per cent. B; the middle 40 per cent. C; the next 20 per cent. D; and the lowest 10 per cent. E. A given judge's ratings are translated over into these new letter units. A rating of A is now always comparable from one rater to another, as in all cases it represents placement in the highest 10 per cent. of the judge's ratings. If Judge X rates every one high, and Judge Z rates every one low, a rating on a seven-point scale of six by X and of three by Z may in both instances receive the same rating. In like manner, ratings of B, C, D and E are comparable;

they occupy the same relative positions in the scales of all judges, thus leveling out differences in judgment standards.

## 2. The Reliability of the Rating Scale

The reliability of a rating scale may be considered from two points of view. According to the first, reliability is measured by the consistency of a given judge's ratings, that is, by the correlation between two sets of ratings made by the same person, but separated by an interval of time sufficiently long to preclude a direct memory effect. According to the second view, the reliability of a judge's ratings is best shown by their correlation with the ratings made by another equally competent judge. Of these alternatives, the second is to be preferred. To be sure, high correlation between two sets of ratings made by the same person tells us that this individual's opinions are stable, and hence worthy of consideration. But close agreement shown by a high correlation between the ratings of two or more competent judges indicates, in addition, a constancy of the traits themselves, rather than a constancy of judgment in a single rater alone. This, of course, is the kind of reliability which one wishes to obtain in ratings.

In general, the reliabilities of rating scales are lower than those of standard intelligence tests. Rugg (51) who has made the most thorough study of the Army man-to-man rating scale reports results with this method which are far from favorable. Working under carefully controlled conditions, and with officers making the judgments who were fully instructed as to the use of the scale, and well acquainted with the ratees, Rugg found that the P.E. of an average rating was five to six points on the eighty-point scale. This result made it exceedingly doubtful whether a single rating would place an individual even within the fifth of the scale wherein he was placed by the average rating or consensus. Rugg is convinced, however, that character traits may be successfully rated (1) if each final rating is the average of three independent ratings, each made on an objectified scale; (2) if the rating scales are comparable and equivalent, and the raters understand and agree upon the meaning of each quality or trait; (3) if the raters are thoroughly acquainted with the person to be rated. Rugg's careful work with the Army rating scale did much to discredit the method of man-to-man rating, and to stimulate work upon other methods.

Studies with the graphic rating scale have, on the whole, been more favorable. Kornhauser (36), using a graphic scale of five intervals, obtained reliability coefficients, when three independent ratings were taken against three other independent ratings, averaging .44. These ratings were made by seven to ten instructors upon 105 students, the traits studied being intelligence, industry, accuracy, cooperativeness, moral trustworthiness and leadership ability. The most reliably rated trait was industry, the average correlation of three judges against three being .58. Hartshorne and May (26) report a constancy of rating on the graphic scale, represented by an average  $r$  of .69, when ratings were made on two occasions by the same teacher. The subjects were 322 pupils in the elementary grades, and the ratings were made by eighteen teachers for "general honesty." The high average reliability of these honesty ratings arises in part from the fact that all scales were discarded upon which not more than two divisions were employed. An average  $r$  of .57 was obtained between the honesty ratings by the classroom teacher and the average ratings of two other teachers, the range being from .37 to .82.

On the Scott Graphic Rating Scale for Workers, Paterson (42) reports an average correlation of .87 between ratings from the second to the third month on the same workers, made by nine foremen. The intercorrelations of the ratings on the same men by seven pairs of foremen averaged .71, for ratings made during the same month. Bradshaw (7) has studied the reliability of the American Council Rating Scale under various conditions. This is a graphic scale consisting of five characteristics, each to be rated on a straight line (p. 106). Correlating the average rating of several judges against a like average of several other judges was found to improve considerably the reliability of the rating. For example, the reliability of the ratings made by three judges on 107 freshmen, *i.e.*, the  $r$  of three ratings against three equally good ratings as predicted by the Spearman-Brown formula, varied from .35 for "appearance" to .73 for "industry." When Bradshaw matched the ratings of ten fraternity men against ten, and twelve against twelve, all ratings being made on each other, the reliability coefficients for all five traits were .90 or above (except for Emotionality, where the  $r$  was .68 in ten against ten). Shen (53) has computed reliability coefficients for ratings on eight traits made by thirteen judges, the number of subjects being 28. The averages of these  $r$ 's range from .34 for "impulsive-

ness" to .71 for "scholarship," the average for all eight being .55. Symonds (60), who has summarized most of the relevant data on rating methods, gives .55 as the typical reliability coefficient for ratings on personality traits. This presumably represents the agreement between two comparable judges.

The reliability of the rating scale, while below that of the standard intelligence test, is still high enough to augur well for the method. Especially is this true when the ratings of several judges are combined. Rating methods will continue to be used by psychologists until objective tests of personality are devised; and, of course, they must be used indefinitely, if such tests are shown to be impracticable.

### 3. The Validity of Rating Scale Estimates

The validity of a set of ratings, like their reliability, may be considered from two points of view. If there is a high intercorrelation among the ratings made by different judges, the averages of such ratings may be taken as valid measures, on the ground that the scale has focused attention upon definite and uniformly recognized characteristics. The average of a large number of independent ratings may also be taken as a criterion, and the validity of a single judge's ratings determined by their correlation with this criterion. Validity may also be defined by the correlation of a set of ratings with objective criteria, such as test results or other measures. This last, of course, is the commonly accepted idea of validity, as it involves the verification of one set of estimates against outside standards. There is much to be said, however, for the view that validity is best expressed by a high correlation among judges, although such relation is often taken to be a measure of reliability rather than of validity (p. 114). A rating scale, however, is a valid and objective measure only to the extent that it is a reliable measure. There is considerable force to the contention that for practical purposes a man possesses intellect, tact or judgment in business matters, if his acquaintances agree that he does. Certainly there is no higher court of appeal. In judging personality traits, therefore, reliability would seem to imply validity.

When all of the factors which reduce the accuracy of ratings are considered, validity correlations between ratings and objective criteria, or among the ratings themselves, are fairly satisfactory. A few typical results will be presented. The correlation between tests

of "self-control" and ratings for the same trait made by their teachers upon 900 public school children has been reported by Harts-horne, May and Maller (27) to be .52. The same authors show that it is possible to increase considerably the correlations between ratings and other measures, when a large number of ratings are combined. Bradshaw (7) obtained an average  $r$  of .66 between composite ratings on the American Council Rating Scale, for twenty-two fraternity men, and a leadership score obtained by rating the various activities in which these men were engaged. The correlation between composite personality ratings made by three raters and grades secured in the first semester was .30. The individual correlations ranged from .50 for grades and industry, to .03 for grades and leadership. The subjects were 107 college freshmen. The correlation of .30 between personality ratings and grades is not a great deal lower than the correlation between intelligence tests and school grades, and indicates the importance of non-intellectual traits in academic achievement. A detailed summary of experimental work upon the validity of rating scales will be found in Symonds (60).

#### 4. Factors Affecting the Accuracy of Ratings

In this section we shall summarize briefly the more important factors which affect the reliability as well as the validity of ratings.

- (a) Close associates will probably rate more reliably than casual acquaintances, but there is little correlation between degree of acquaintance and competence as a rater.
- (b) Raters tend to rate their friends too high on desirable traits and too low on undesirable.
- (c) Individuals differ markedly in their ability to make judgments. Ratings of which the judge expresses himself as very sure are in general more reliable than ordinary ratings.
- (d) Characteristics or traits exhibited in one's reactions to assigned tasks, or to outside situations are better rated than social or personal traits which are difficult to observe. Scholarship, industry, leadership, for example, are better rated than impulsiveness, adaptability, or common sense. Ratings are often more reliable when a general characteristic is broken up into several constituent parts.
- (e) There is a decided tendency for a rater to over-rate himself

on desirable and under-rate himself on undesirable traits as judged by the consensus.

- (f) "Halo effect." The general tendency to rate an individual who has been placed high or low on a general trait, *e.g.*, intelligence, consistently high or low on a number of more specific traits, is called the "halo effect." It represents a general carry-over, a "tendency to think of a person in general as rather good or rather inferior, and to color the judgments of the qualities by this general feeling" (63). Thorndike notes further (63) that "general merit as a teacher has correlations of .68 with intellect, .79 with power in discipline and .63 with voice. It is clear that the rating of a teacher's voice must have been influenced by the general impression of her ability. Voice correlates .50 with 'interest in community affairs' and .63 with intellect!" The sample in this study quoted by Thorndike consisted of 129 teachers.

The remedy for the halo effect is to have all the members of the group rated for one trait at a time, instead of having each person rated on all traits at once. The halo effect is lessened when the rater is cautioned against it, and when the traits to be rated are objectively and precisely defined.

### REPRESENTATIVE RATING SCALES

1. Behavior Rating Schedules, by M. E. Haggerty, W. C. Olson, and E. K. Wickman

*Date:* 1930.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Pre-school and elementary school children.

*Contents:* The behavior rating schedule consists of two parts, A and B. Schedule A presents a list of behavior problems, the occurrence or non-occurrence of which is to be checked by the rater. Schedule B is a graphic rating scale which contains thirty-five intellectual, physical, social and emotional traits: Samples:

Schedule A:

Cheating      0      4      6      7

Schedule B:

15. Is he quiet or talkative?

*Scoring:* In Schedule A the degree to which a child possesses a given trait determines his score. Total score is the sum of the individual ratings. In Schedule B, graphic ratings are translated into numerical values and totaled to give the final rating.

*Norms:* Schedule A norms are based upon 2,163 children; Schedule B norms upon 2,867. Norms for an abbreviated form of Schedule B, used with pre-school children, are based upon ninety cases.

*Reliability:* .66 for pre-school children; .86 for elementary school children, repeated ratings by same teacher. N not reported. Validity .60 between Schedules A and B.

2. Diagnostic Analysis of Effective Leadership, Personal Inventory D 1, by Donald Laird

*Date:* 1929.

*Publisher:* Hamilton Republican, Hamilton, New York.

*Designed for:* Adults, especially those in industry.

*Contents:* Twelve traits are to be rated, which are presumably important in leadership, *viz.*, self-confidence, organizing ability, commercial attitudes, constructive thinking, *etc.* Questions concerning the possession or non-possession of specific qualities under each general trait are to be answered by checking in the Yes or No column; and a general rating is given on a graphic scale.

*Scoring:* Checks on a graphic scale are translated into numerical values and totaled to give a composite rating on the twelve traits. Scores on specific elements in, or aspects of, each general trait may be determined from a key.

*Norms:* Average total scores for strong and weak executives are given.

*Reliability:* Not reported.

3. Graphic Rating Report on Workers, Scale B, by the Scott Company

*Date:* 1922.

*Publisher:* Out of print.

*Contents:* Seven qualities to be rated on a graphic scale. These qualities are ability to learn, quantity of work, quality of work, industry, initiative, coöperativeness, knowledge of work.

*Scoring:* Ratings are converted into numerical values on a scale from 1 to 10.

*Norms:* None reported.

*Reliability:* Average *r* of .76 between ratings given in first and second months by nine foremen. The *r*'s between the ratings on the same workers made by three pairs of foremen for the third month vary from .50 to .90.

*Reference:* PATERSON, D. G., "The Scott Company Graphic Rating Scale," *Journal Personnel Research*, 1:361-376, 1922.

4. North Carolina Rating Scale for Fundamental Traits, by Floyd H. Allport

*Date:* 1924.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Designed for:* Adults, primarily; can be used with adolescents.

*Contents:* Twenty-four paired statements expressing the extremes of a given quality or trait. Ratings are made on a nine-point scale. Example:



5. Slow in decision and action.....1 2 3 4 5 6 7 8 9.....quick in decision and action.

*Scoring:* Subject is rated by marking through the number which indicates the degree of the trait possessed.

*Norms:* None reported.

*Reliability:* Not reported.

5. Occupational Intelligence Scale, by F. E. Barr

*Date:* 1918.

*Designed for:* Adults.

*Contents:* 100 representative occupations are listed, each with a numerical value which designates the amount of intelligence presumably required for it. The method of construction of this scale is given in the Terman reference below. Samples:

3.62 Day laborer

9.37 Carpenter

15.75 Lawyer

*Scoring:* The individual's intelligence rating is determined from his occupational status.

*Norms:* None reported.

*Reliability:* Not reported.

*Reference:* TERMAN, L. M., *Genetic Studies of Genius*, I, 66-72, 1925.

6. Personality Rating Scale, Committee on Personality Measurement, American Council on Education

*Date:* 1928. Revisions A and B, 1929.

*Publisher:* American Council on Education, 744 Jackson Place, Washington, D. C.

*Designed for:* Students, primarily college freshmen.

*Contents:* This scale consists of five traits to be rated on a graphic scale. The traits are (1) Appearance and manner; (2) industry; (3) ability to control others; (4) emotional control; (5) distribution of time and energy. Space is provided on this scale for "behaviorgrams," i.e., concrete instances in support of the rater's judgment.

*Scoring:* Revision A is scored by checking along a straight line which is subdivided into ten sections. Revision B is scored by checking one of six alternatives.

*Reliability:* .77 for the average of three raters against three other raters. N = 107 freshmen.

7. Rating Scale for Teachers, by H. C. Almy and Herbert Sorenson

*Date:* 1930.

*Publisher:* Public School Publishing Co., Bloomington, Illinois.

*Designed for:* Teachers.

*Contents:* Twenty traits to be rated on a graphic scale. These traits are resourcefulness, tact, fairness, sympathy, patience, foresight, etc. A graphic scale for indicating whether the basis of the rater's judgment is definite, general or inadequate is provided.

*Scoring:* The rating line is subdivided into ten equal parts. The score is the sum of the points earned on each of the traits.

*Reliability:* .92, given by the authors for first and second ratings, same judge. N = 110 practice teachers.

8. **Sims' Score Card for Socio-Economic Status**, by V. M. Sims

*Date:* 1927.

*Publisher:* Public School Publishing Company, Bloomington, Illinois.

*Designed for:* Children in Grades 4 to 12; can be given individually or in groups.

*Contents:* The purpose of the Score Card is to secure information concerning the social, economic and cultural status of a home. Sample questions are:

1. Have you a telephone in your home? Yes No
5. Did your father go to college? Yes No
15. Does your family attend concerts? Never Occasionally Frequently
20. How many magazines are regularly taken in your home?  
None One Two Three or More
21. About how many books are in your home? None 1 to 25 26 to 125 126 to 500 More
22. How many rooms does your family occupy? 2 3 4 5 6 7 8 9 10 11 12 More
23. Write your father's occupation on this line.....

*Scoring:* Point credits ranging from 0 to 6 are assigned the answers. The method of combining the separate score values is given in the Manual.

*Norms:* Percentile norms based on 686 sixth-, seventh- and eighth-grade children.

*Reliability:* .91 (split-half) for 686 children.

*Reference:* SIMS, V. M., *The Measurement of Socio-Economic Status*, Public School Publishing Company, Bloomington, Illinois, 1928.

9. **Whittier Home Rating Scale**, by J. H. Williams

*Date:* 1916.

*Publisher:* Whittier State School, Whittier, California.

*Designed for:* Studying the cultural and economic level of the home.

*Contents:* This scale contains a score card with directions for grading each of five different items: necessities, e.g., furnishings, in the home; neatness; size of the home, number of rooms; parental conditions, e.g., parents living together, mother dead; parental supervision.

*Scoring:* Each item is rated on a scale from 1 to 5. Concrete illustrations of what is meant by each grade are given on the score card. From these a total Index is calculated.

*Norms:* Ratings for fifty homes of non-delinquent children, and 120

homes of delinquents are given. Ratings for gifted children have been given by Terman (62).

*Reliability:* Not given.

*Reference:* WILLIAMS, J. HAROLD, "The Whittier Scale for Grading Home Conditions," *Journal Delinquency*, 1:273-286, 1916.

### THE QUESTIONNAIRE

The questionnaire and the rating scale overlap in certain important respects which will be noted later. Both are concerned primarily with traits not open to direct measurement. They differ chiefly in that the rating scale consists of quantitative estimates or judgments, while the questionnaire presents a systematic report of an individual's thoughts, attitudes, or experiences. When a person, in answering a questionnaire, gives estimates of the degree to which he possesses certain traits or characteristics; or when, in filling out a rating scale, questions of attitude or questions of fact are involved, the two techniques are essentially the same.

As distinguished from the psychological test, which records what a person can do when given a definite task to perform, the questionnaire presents a personal record of an individual's social or moral attitudes, his interests, his beliefs, and his convictions. The person who answers a questionnaire is not scored in terms of time taken to complete, or amount done. Rather, his score serves to identify him, temperamentally or socially, with a more or less well-defined group which thinks much as he does or believes and is interested in the same things.

In general, questionnaires have been used by psychologists for three purposes. The first is to secure data concerning an individual's so-called maladjustments, his feelings of inadequacy, fears, worries and the like. Such "personal data sheets" are, in effect, records of nervous and mental symptoms. Secondly, questionnaires have been used in studies of personality in order to obtain attitudes or beliefs upon social, economic and religious issues. Thirdly, questionnaires are employed to discover systematic interests in people, books, sports, vocations, mechanical and social activities, and the like. In addition to these various uses, it should be noted that questionnaires have also been widely used by psychologists and sociologists to obtain data on home conditions, occupational status, cultural level, and other environmental facts.

## THE CONSTRUCTION AND USE OF QUESTIONNAIRES

## 1. Selection of Items

In general, the person constructing a questionnaire first makes an analysis of the characteristics he wishes to measure, drawing his material from the work of recognized authorities in the field of his interest. Woodworth (32), in constructing his P.D.<sup>1</sup> Sheet, for example, assembled some 200 questions taken from authorities in psychiatry, and dealing with symptoms found generally to antedate nervous and mental breakdowns. Freyd's (19) list of fifty-four characteristics of introversion, later used by Heiddreder in the form of a questionnaire, is a compilation of the opinions of authorities upon this topic. Allport's study of ascendance-submission, Watson's study of fair-mindedness, Strong's study of vocational interests, to mention only a few, represent in each case a consensus of expert opinion upon the topics under investigation (p. 131).

An empirical method of selecting the items of a questionnaire is often used in conjunction with the method just mentioned. This is the method of *experimental tryout*, in which the final items of the questionnaire are selected as a result of actual trial. In Woodworth's P.D. Sheet (32), for instance, the questions were submitted to groups of college students and to nearly 1,000 drafted men—all presumably normal. The questions were also given to men diagnosed as psychoneurotic. Two criteria were set up which the individual items were required to meet before inclusion in the final form of the questionnaire. First, all questions which returned a large percentage of unfavorable answers (greater than 25 per cent.) in the normal group were omitted or "stiffened," on the ground that a symptom which appeared so frequently in normal people could not be taken to indicate abnormality. Secondly, a 2:1 ratio was adopted as between psychoneurotic and normal responses. That is, a symptom had to be reported twice as often by psychoneurotics as by normals to be taken as indicative of an abnormal trend. In House's revision (32) of the Woodworth P.D. Sheet a 3:2 ratio as between psychoneurotic and normal responses was employed as well as a 2:1 ratio, in order to increase the number of questions.

Akin to this method "of discriminative ratios" just described is

<sup>1</sup> Personal data.

the criterion of *internal consistency* used by the Thurstones (67). The Thurstones' final form of their Neurotic Inventory represents a selection of 223 items, out of more than 600, taken from the Woodworth P.D. Sheet, and the lists of Freyd, Laird, and Allport. On the basis of the total scores made by 694 college students on the Neurotic Inventory, a group of fifty students was cut off from each of the two extremes (high and low ends) of the distribution. Group 1, those fifty students having the lowest total scores, was taken to constitute the probably best-adjusted group; Group 2, those fifty students having the highest total scores, the probably most-neurotic group. Only those questions were retained in the Inventory which exhibited a significantly greater percentage of unfavorable answers in Group 2 than in Group 1. The criterion of internal consistency is, as its name implies, a method of securing homogeneity within a questionnaire, *i.e.*, of securing items which will tap correlated behavior activities. As a means of validation it does not, of course, go beyond the material in hand.

In studying introversion, Heidbreder (29) employed the empirical method in evaluating her questionnaire. Freyd's fifty-four items (19), covering the characteristics most often attributed to introverts, were presented to 200 men and 100 women students, with the request that they rate themselves + if the characteristic applied, — if it did not apply, and ? if the rater was doubtful. Each of the persons in this group had been rated on the fifty-four items by two associates (as well as by himself) so that three ratings in all were available for each subject. The score on each paper consisted of the algebraic sum of the plus and minus items—a plus score indicating introversion and a minus score extroversion.

Heidbreder's distribution of combined self and associate ratings for her 200 subjects was approximately normal, although the distribution was somewhat shifted over in the direction of extroversion, the mean score being —11.25. Neither introversion nor extroversion, therefore, but "ambiversion" (or a mild extroversion) was the typical attitude. In order to test the diagnostic value of each item, Heidbreder selected for comparison, on the basis of total scores, the 25 per cent. most introverted students and the 25 per cent. least introverted students. When the difference in the percentage marking an item introvert was reliably greater in the introvert group than in

the extrovert group, the item was taken to be diagnostic of introversion.<sup>1</sup>

Ream (48) was one of the first investigators to use the method of contrasted groups as a means of finding significant differences in expressed interests. Ream compared two groups of salesmen in order to discover whether the successful could be distinguished from the unsuccessful by their likes and dislikes. Whenever the difference in the percentage of successful and unsuccessful salesmen marking an item L (like), I (indifferent), or D (dislike) equaled the S.D. (diff.), the item was considered to have distinguished adequately between the preferences of the two groups.<sup>1</sup> Using a revision and selection from the 1921 Carnegie Interest Inventory, Freyd (20) employed the empirical method in selecting those activities and interests which are peculiar to the "mechanically minded" individual. The first of Freyd's lists contained seventy-two occupations, covering various kinds of work activities; the other 129 items described characteristics of people, as well as a variety of activities and situations, which people generally like or dislike. Likes and dislikes in varying degrees were recorded by circling one of five symbols: L! L ? D D! In order to determine whether an item was more often preferred by the "mechanically minded" group (twenty-nine engineers) than by the "socially minded" group (thirty salesmen), Freyd calculated the percentages of each group, marking the item liked or disliked in varying degree. A  $\frac{\text{percentage difference}}{\text{S.D. (diff.)}}$

greater than 2 for any item was taken to indicate a definite preference for that item by one of the two groups.<sup>1</sup>

Hubbard's (33) Interest Analysis Blank is another illustration of the use of contrasted groups in the selection of the items of a questionnaire. Hubbard's blank consists in part of a list of 115 occupations each followed by the letters L D O U. If the subject crosses out the L it means that he likes the occupation named; if he crosses out the D it means that he dislikes it; if he crosses out the O it means that he has no particular feeling either way; and if he crosses out the U it means that he knows nothing about the occupation in question. Likes and dislikes for a series of sixty-three activi-

<sup>1</sup>For the method of calculating the significance of a percentage difference, see Holzinger, K. J., *Statistical Methods for Students in Education*, Ginn & Co., Boston, chapter 13, 1928.

ties were indicated in the same way. In devising a scoring scheme for the blank, two groups of boys were selected, the one group possessing definite mechanical abilities, the other group having little mechanical ability. When the percentage of the mechanically inclined group "liking" an item exceeded the percentage of the non-mechanically inclined, the item was considered to be indicative of "mechanical interest." Plus scores were assigned to those items which revealed the interests of the mechanically gifted boys, and minus scores to items which characterized the interests of the non-mechanically inclined boys. The total score, which was the algebraic sum of these plus and minus scores, thus served to identify the boy with one or the other group.

## 2. Scoring Methods and Answer Techniques

The simplest scoring technique for the items of a questionnaire is illustrated by the Woodworth P.D. Sheet in which each question is answered by circling the Yes or No which follows it. House (32), in his revision of the P.D. Sheet, increased the range of answers by allowing Yes to be answered as "Extreme" or "Moderate." The Thurstones (68) and Bernreuter (4) allowed three choices in their schedules, Yes, No and ?, the last meaning undecided. Ream used three categories, L I D; Freyd five categories, L! L ? D D! G. B. Watson (71) has used a five-fold multiple-choice method in recording answers to parts of his test, *viz.*, +2 +1 0 -1 -2; or All, Most, Many, Few, No.

When the behavior described seems to fall naturally into two opposing categories, contrasted statements expressing the extremes of a given activity are often employed in recording answers. Marston (38), in studying tendencies to introversion and extroversion in young children, required his judges to check that one of two contrasting statements which better applied to the child being described. When two forms of behavior are thus placed in opposition the contrast is better brought out, since the judge has before him the extremes of the given activity. Cady (9) has used an interesting variation of this method in checking the reliability of the answers to his revision of the P.D. Sheet. Each question in his original list was rephrased in a second form so as to require an opposite response, thus supplying a check upon the consistency of the subject's answers. Pressey (46) has checked upon the subject's attitude

in his X-O tests by including a single innocuous word, or "joker," in each list of five words (p. 136). A notion of the subject's attitude toward the test, whether he takes it seriously or is flippant and careless, is obtained by noting the number of "jokers" which are crossed out.

G. B. Watson (71) has employed several ingenious methods in his study of "fair-mindedness" or lack of prejudice. To illustrate, in Form C "Inference Test" and in Form E "Arguments Test," the assumption is made that biased or prejudiced views will lead one to draw extreme conclusions which are not logically valid. An extreme inference or argument, therefore, whether *pro* or *con* is taken as evidence of prejudice. Form D "Moral Judgments" is based upon the assumption that individuals are more likely to be governed by prejudice and preconceived notions when the situation to be evaluated is personal and immediate than when it is remote and detached. Hence, paired situations are presented, the one referring to events historically or geographically remote, the other to happenings which are more immediate and personal. The same principle, however, underlies both situations. It is assumed that prejudice will be shown by inconsistency in checking arguments relating on the one hand to remote, and on the other to immediate, situations.

Recourse may be had to a consensus of expert opinion (a) in selecting the items of a questionnaire, and (b) in determining just what the attitude expressed by the answer probably denotes. Symonds (61), in his "Social Attitudes Questionnaire," decided upon the liberal as opposed to the conservative side of questions dealing with social issues, by having five experts designate what they considered to be the liberal, and what the conservative, position. In the same way, Hart (25) used a group of "leaders of social progress" to determine the liberal or socialized answers to the questions in a test dealing with social attitudes and interests.

The Colgate Mental Hygiene Tests (37) introduce the graphic rating scale technique into the scoring of the questionnaire. The answer to each question on these tests is given by checking along a line which consists of ten one-half-inch segments. Scoring is accomplished by means of transparent stencils. These stencils contain lines of varying length, each of which corresponds to that interval upon a line within which the answer to a question is to be judged unfavorable. Stencil lines fit over the test page and are coincident



with the rating lines. The extent of the unfavorable interval assigned to each question was determined by throwing into a distribution the answers (expressed as linear magnitudes) given to each question by about 2,000 subjects. The position of the first quartile from the "neurotic" end of the distribution was then calculated, and any answer falling within this section was judged to be unfavorable. In other words, an unfavorable answer places one in the lowest 25 per cent. of individuals at the "neurotic" end of the distribution.

### 3. Weighting and Scaling Methods

The use of "contrasted groups" in determining the relative strength of preferences has already been touched upon briefly (p. 125). The difference between the percentages of two such groups marking an item as preferred or "liked" offers, too, a simple means of weighting the item for one or the other group. This method of weighting the answers to a questionnaire is illustrated by the work of Garretson (24). This investigator studied the interests of three groups of boys enrolled, respectively, in academic, commercial, and technical high schools, with a view toward vocational guidance. Garretson's questionnaire consisted of 328 items, covering occupations and other activities, school subjects and interests, celebrities, things a boy would like to own, magazines read, *etc.* Answers were indicated by circling L I D. This questionnaire was administered to 483 boys in a commercial high school; to 503 boys in a regular academic high school; and to 596 boys in a technical high school. All of these boys were in the first year of their high-school courses.

Weights or numerical values were assigned to the answers given to each item in the questionnaire, according to whether they denoted technical, academic, or commercial interests. These weights were determined by dividing the difference in the percentages of two groups marking a given item by the S.D. of this difference.<sup>1</sup> The nearest whole number was taken as the score. To illustrate, if the boys in the technical high school indicated a preference for "electric motors" to the extent of 10 per cent. over the boys in the academic high school, and if the S.D. (diff.) equaled 5 per cent., the weight of this item was 2 on the scale of technical interests. The groups employed in determining item weights contained approximately 150 boys, of average ability in their studies, selected from the three schools in which the tests were given.

<sup>1</sup> See p. 125, footnote.

Garretson has checked his scoring method against other more elaborate weighting systems, *e.g.*, that of Strong. His results indicated that the simple method gave as high validity correlations with his criteria as did more involved methods. In his revised questionnaire, Garretson has introduced further simplification by using only +, 0, and — as scoring designations. A + item has a weight of 2, a 0 (neutral) item a weight of 1, and a — item a weight of 0.

The Allport A-S<sup>1</sup> Study offers a good illustration of item weighting, when several possible answers are allowed. This questionnaire (p. 131) contains in its Form for Men, thirty-three situations, and in its Form for Women, thirty-five situations, to each of which two to five different responses are suggested. The numerical score value to be attached to each response was determined in the following manner: Five ratings for ascendance-submission, *viz.*, one self-rating and four ratings by associates, were obtained upon 400 men and 200 women—all college students. A seven-point scale was used for these ratings, 1 meaning most ascendant, and 7 most submissive. The averages of these five ratings constituted the criterion, and were found, when put into a frequency distribution, to exhibit a surprising degree of normality. The range of average ratings, for instance, was from 1.4 to 6, the mean being 3.48. The Allports' first method of obtaining the various weights to be attached to the different alternative responses to the questionnaire was to compute the average criterion scale rating of all subjects answering an item in a given way. For example, if the average criterion rating of all of those who gave the response "habitually" to situation 1 (p. 132) was 4, then 4 would be the weight assigned to this answer. An alternative and simpler weighting scheme was later used. This consisted in subtracting the average criterion scale rating for each response from the mean criterion scale rating, *i.e.*, 3.48. For example, if the response "never" had a mean criterion rating of 5.68, then 3.48 — 5.68 gives —2.20, as the weight to be assigned to this item. Minus signs are taken to indicate submissive, plus signs ascendant, behavior.

The Strong Vocational Interest Blank (55, 58) represents a considerable expansion and revision of the interest Analysis Blank used earlier by Cowdery (13). The contents of the Vocational Interest Blank are described on p. 143. This questionnaire is scored sepa-

<sup>1</sup> Ascendance-submission.

rately to show the interests of each occupation, the differences in the numerical values assigned to the 420 items in the blank denoting the variation in degree of interest of men found in different occupations. Scoring stencils have been developed for many occupational groups, ranging in interests and intellectual requirements from advertiser to Y.M.C.A. physical director.

If the experimenter wishes to know whether or not his subject has the interests of a lawyer, a scoring stencil with one set of weights is used; if he wishes to know whether the subject's interests are those of a chemist or architect, other stencils, in which the items are differently weighted, are employed. Strong's method of computing the numerical weights to be attached to the items of his Vocational Interest Blank, when scored, let us say, for "engineer" interest, was to compare the answers of a group of successful engineers with the answers of "men in general." The greater the proportion of engineers liking a certain vocation, a certain sport, or type of reading matter, over men in general, the greater the weight assigned to this item on the engineer interest scale. For example, the answers given by 575 engineers and 3,920 men not engaged in engineering ("men in general") to the item Architect were distributed in percentages as follows:

	Engineers	Men in General
L .....	57%	42%
I .....	32%	37%
D.....	11%	21%

It is evident that "liking to be an architect" is characteristic of engineers to a somewhat greater degree than of men in general, and hence this vocation has positive weight on a scale of engineer interest. The weights assigned to the other items on the engineer interest scale were calculated by Strong from formulas adapted from Cowdery (13).

Strong (56) has made a thorough investigation of the relation between weighted and unweighted scoring keys. In three groups, lawyers, architects, and certified public accountants, the size of the groups varying from forty-five to 190, he obtained an average correlation between weighted and unweighted items of .93. This finding would seem to indicate that weighted answers are of no greater value than unweighted. In a comparison of individual scores for four different occupations, however, Strong found that the weighted

keys classified nearly twice as many men as *not* belonging to some other than their own occupation, as the unweighted keys. Expressed more concretely, when weighted scoring keys were employed, fewer architects were rated as having the interests of a lawyer, and fewer journalists as having the interests of accountants. Because of this more clear-cut separation of occupational groups as regards their interests, Strong inclines to the use of a weighted scoring method.

Thurstone's contributions to the construction of attitude scales are too extensive both in technique and method for more than a cursory description here. Thurstone's work, which is based upon the psychophysical methods, leads to scales in which each item, or each situation, is placed by the consensus of judgment along an attitude continuum. The zero, or reference point, on each scale is known, and the distance of each item from the next is determined with reference to this starting point. To illustrate, in a study of nationality preferences, Thurstone (64) constructed a scale in which twenty-one nationalities were arranged in order, in terms of the willingness of Americans to associate with them. Each group was compared with every other, *e.g.*, Americans with Hindus, Japs with Italians, English with Swedes, the judges being 239 undergraduates. The percentage of preference judgments was determined for all possible pairs, and from these the scale value of each group was calculated. In the final preference scale, Americans were taken as the zero or reference point since all other nationalities were less often preferred. Englishmen and Scotchmen rank next in order, Turks and Negroes next to last and last, respectively. Many other scales have been constructed by Thurstone dealing with attitudes toward prohibition, the movies, divorce, the Bible, Communism, *etc.* Thurstone's fundamental techniques, used in attitude measurement, are contained in several articles which should be consulted by the student (65, 66).

## REPRESENTATIVE QUESTIONNAIRES

### I. PERSONALITY INVENTORIES AND ADJUSTMENT QUESTIONNAIRES

1. A Scale for Measuring Ascendancy-Submission in Personality (The A-S Reaction Study), by G. W. and F. H. Allport

*Date:* 1928.

*Publisher:* Houghton Mifflin Company, New York.

*Purpose:* To "discover the disposition of an individual to dominate his

fellows (or to be dominated by them) in various face-to-face relationships of every-day life."

*Designed for:* Men and women, a separate form for each.

*Contents:* The A-S questionnaire, Form for Men, comprises thirty-three, and the Form for Women thirty-five, situations or problems. To each of these, two to five alternative responses are given. Some situations appear on only the one form, others on both forms. Samples are:

*For both sexes:*

At church, a lecture, or entertainment, if you arrive after the program has commenced and find that there are people standing, but also that there are front seats available which might be secured without "piggishness" or discourtesy, but with considerable conspicuousness, do you take the seats?

habitually .....  
occasionally .....  
never .....

*For men:*

In witnessing a game of football or baseball in a crowd have you intentionally made remarks (witty, encouraging, disparaging, or otherwise) which were clearly audible to those around you?

frequently .....  
occasionally .....  
never .....

*For women:*

If you made purchases at Woolworth's or at the bargain counter, would you mind your friends' knowing it?

sometimes .....  
no .....

*Scoring:* Numerical values, plus, minus, and zero, are assigned to the two to five alternative answers to each situation. Separate scoring values are supplied for men and women. The final score is the algebraic sum of the scores on the separate items. This total is translated into a rating which gives the degree of ascendance or submission.

*Norms:* Tentative norms are given in the Manual, based upon groups of 1,860 men and 1,275 women.

*Reliability:* .74 (split-half) for 400 men; .78 (retest) for 200 women.

## 2. Colgate Mental Hygiene Tests, Personal Inventory B2, by Donald Laird

*Date:* 1925.

*Publisher:* Hamilton Republican, Hamilton, New York.

*Purpose:* To detect abnormal social and emotional trends in adults.

*Designed for:* Adults; used mostly with college students.

*Contents:* A total of sixty-six questions, each to be answered on a graphic scale. This scale consists of lines which have been subdivided into ten one-half-inch segments. Descriptive terms are printed beneath

each line. The questions in the inventory are divided into three categories: Part I, thirty-two questions dealing with psychasthenia; Part II, fourteen questions dealing with schizophrenia; Part III, twenty items dealing with neurasthenia.

*Scoring:* The subject answers a question by checking along the line. If the check falls in the "neurotic" section of the line (given in the key), the answer is scored as unfavorable. Final score is the sum of unfavorable answers.

*Norms:* Percentile norms are given for college men and women, in the Manual accompanying the test.

*Reliability:* Laird gives .88 (N not given). Flemming (16) reports a reliability coefficient of .78 (split-half) in a group of 332 college freshmen.

*Reference:* LAIRD, D., "Detecting Abnormal Behavior," *Journal Abnormal Psychology*, 20:128-141, 1925.

### 3. Emotional Maturity (E.M.) Scale, by R. R. Willoughby

*Date:* 1931.

*Publisher:* Stanford University Press, Stanford, California.

*Purpose:* To estimate emotional maturity as shown by subject's willingness to accept responsibility, freedom from infantile motives, etc.

*Designed for:* College students and adults.

*Contents:* Sixty items, "each of which describes in terms of a hypothetical subject (S) a type of situation and a reaction to it." Examples are:

- 2. S is extremely solicitous of his immediate family associates.....
- 15. S demands that he be punctiliously served in hotels, sleeping cars, etc.....
- 58. S is serious and anxious in his manner, even in cases where nothing important can hang on the results.....

*Scoring:* The person taking the test checks a statement if it describes him, or if it describes the person whose characteristics he is evaluating.

*Norms:* Percentile norms for seventy students, of both sexes, each rated by self and by two acquaintances.

*Reliability:* .54 between college student raters. Author believes reliability is much higher with experienced raters.

### 4. "Experience Variables" Record, by J. O. Chassell

*Date:* 1928.

*Published:* By author, University of Rochester Medical School, Rochester, New York.

*Purpose:* To provide a systematic survey of an individual's attitudes toward his family; sex and religious standards; social and emotional adjustments; intellectual and vocational interests and aptitudes.

*Designed for:* Adults.

*Contents:* There are twelve sections. Within each section questions are

classified into four groups, in accordance with the kind of behavior investigated.

*Scoring:* Time of appearance of a given experience is recorded by checking in one of three columns, headed "Childhood," "Early Teens," "Recent or New," respectively. Record is used as a case history. Author advises use of record in personal counseling, especially in interviews with college students.

*Norms:* None reported.

*Reliability:* Reliability coefficients for different groups of items are reported by Chassell, in terms of C, coefficient of contingency.

*Reference:* CHASSELL, J. O., *Experience Variables, A Study of the Variable Factors in Experience Contributing to the Formation of Personality*, published privately, Rochester, N. Y., 1928.

5. Mental Hygiene Inventory (a Revision of the Woodworth P.D. Sheet) by S. D. House

*Date:* 1927.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Purpose:* To measure neurotic tendencies in adults.

*Designed for:* Adults (men).

*Contents:* Seventy-five statements, twenty-five referring to experiences before fourteen years of age; fifty to experiences after fourteen years. Each question is followed by Yes (extreme or moderate) No.

*Scoring:* Yes—extreme, Yes—moderate, or No circled by subject. Score is total number of unfavorable answers.

*Norms:* Results are reported for various college groups and for neurotic patients.

*Reliability:* .71 (retest) in group of sixty-eight college students; .85 (retest) in group of fifty-eight students.

*Reference:* HOUSE, S. D., "Mental Hygiene Inventory," *Archives Psychology*, 88, 1927.

6. Personal Data Sheet, by R. S. Woodworth

*Date:* 1917.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Purpose:* To estimate an individual's psychoneurotic tendencies, i.e., in general, his emotional and social inadequacies.

*Designed for:* Adults; originally for soldiers.

*Contents:* One hundred sixteen questions, each followed by Yes and No. Samples are:

1. Do you usually feel strong and well? Yes No

18. Do you feel tired most of the time? Yes No

35. Were you shy with other boys? Yes No

64. Does it make you uneasy to go into a tunnel or subway? Yes No

97. Do your interests change quickly? Yes No

110. Has any of your family committed suicide? Yes No

*Scoring:* Each question is answered by circling Yes or No. The score is the total number of unfavorable or neurotic answers, the "wrong" answer being sometimes Yes and sometimes No.

*Norms:* An average score of thirty-six for neurotics and a ten for normals is given by Woodworth. For results with other groups see Fleming (16) and Hollingworth (31).

*Reliability:* .90 (split-half), with groups of soldiers. N not reported.

7. Personality Inventory, by R. G. Bernreuter

*Date:* 1931.

*Publisher:* Stanford University Press, Stanford, California.

*Purpose:* To measure (1) neurotic tendencies, (2) self-sufficiency, (3) introversion-extroversion, (4) dominance-submission.

*Designed for:* High schools and college students; also adults.

*Contents:* One hundred twenty-five questions, taken from Thurstone, Laird, Allport, and others. Each question is followed by Yes, No, ?.

*Scoring:* This blank is scored separately upon four scales, a different set of weights being assigned to the items in each case. Scale 1 is for neurotic tendencies; Scale 2 for self-sufficiency; Scale 3 for introversion-extroversion; Scale 4 for dominance-submission.

*Norms:* Percentile norms for high-school and college students, and for adults, both men and women, are supplied for each scale.

*Reliability:* Scale 1, .88; Scale 2, .85; Scale 3, .85; Scale 4, .88. N = 128 college students in each case.

8. Personality Schedule, by L. L. and T. G. Thurstone

*Date:* 1929.

*Publisher:* University of Chicago Press, Chicago, Illinois.

*Purpose:* To detect personal and social maladjustments.

*Designed for:* Adults, especially college students.

*Contents:* Two hundred twenty-three questions followed by Yes, No and ?, the last meaning not sure or undecided.

*Scoring:* Subjects are instructed to circle Yes, No or ?. The maladjustment score is the sum of unfavorable answers. Scoring is facilitated by means of a stencil.

*Norms:* For college students, score ranges corresponding to normal, slightly maladjusted, maladjusted, etc., are given.

*Reliability:* .95 (split-half) for 694 Chicago University freshmen.

*Reference:* THURSTONE, L. L., AND T. G., "A Neurotic Inventory," *Journal Social Psychology*, 1:1-30, 1930.

9. Revision of the Woodworth P.D. Sheet, by Ellen Mathews

*Date:* 1923.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Purpose:* To measure the emotional stability of children.

*Designed for:* Ages eight and above.

*Contents:* Form A, sixty questions; Form B, forty questions, each question followed by Yes or No. These two forms were later combined into



a single form of seventy-five items. This questionnaire is an expurgated and simplified edition of the original Woodworth P.D. Sheet.

*Scoring:* Each question is answered by circling Yes or No. The final score is the total number of unfavorable answers.

*Norms:* Medians, twenty-fifth and seventy-fifth percentiles given by age groups for boys and girls, nine to nineteen years old. Also medians, twenty-fifth and seventy-fifth percentiles for Italian and Jewish children, for retarded and advanced pupils, and for other special groups.

*Reference:* MATHEWS, ELLEN, "A Study of Emotional Stability in Children," *Journal Delinquency*, 8:1-40, 1923.

10. X-O Tests for Investigating the Emotions, by S. L. Pressey

*Date:* 1920.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Purpose:* To uncover individual differences in moral, emotional and affective tendencies; eccentric attitudes, anxieties and other behavior abnormalities.

*Designed for:* Form A, for adults; Form B, for children.

*Contents:* The blank consists of four sub-tests, each test containing twenty-five sets of five words each. In tests 1, 3 and 4 the subject is instructed to cross out all words denoting situations or things which are unpleasant, blameworthy or emotionally disturbing; and to circle the one word in each list of five which is most unpleasant or disturbing. In test 2 the subject is instructed to cross out all words associated in any way with a given stimulus word and to circle the one word in each set having the closest relation to the stimulus word. Samples are:

Test 1. Disgust, fear, sex, suspicion, aunt.

Test 2. Blossom: flame, flower, paralyzed, red, sew.

Test 3. Begging, swearing, smoking, flirting, spitting.

Test 4. Injustice, noise, self-consciousness, discouragement, germs.

There are 600 separate elements in the whole test.

*Scoring:* The total number of words crossed out gives the "total affectivity score," while the sum of the words crossed out by S which are different from the modal word in each line, *i.e.*, the word most often chosen, gives the deviation or "total idiosyncrasy score."

*Norms:* Median and quartile scores both for total affectivity and total idiosyncrasy are given in the Manual. Results are based upon test scores of 114 college students, fifty-eight women and fifty-six men.

*Reliability:* Test 1, .85; Test 2, .86; Test 3, .82; Test 4, .87 by retest on sixty-four college students after a forty-eight-hour interval (40). These reliability coefficients are lower when a longer time interval has elapsed. Flemming (16) reports a reliability coefficient of .97 (split-half) for total affectivity score obtained in a group of 328 freshmen; and a reliability coefficient of .50 (split-half) for the total deviation score in a group of 311 freshmen.

*Reference:* PRESSEY, S. L., "A Group Scale for Investigating the Emotions," *Journal Abnormal and Social Psychology*, 16:55-64, 1921.

## II. INTROVERSION-EXTROVERSION QUESTIONNAIRES

### 1. Colgate Mental Hygiene Tests, Personal Inventory C2, by D. Laird

*Date:* 1925.

*Publisher:* Hamilton Republican, Hamilton, New York.

*Purpose:* To measure tendencies toward introversion.

*Designed for:* Adults, used mostly with college students.

*Contents:* Forty-eight items to be answered on a graphic scale in terms of the behavior habitually shown by the subject.

*Scoring:* Answers are checked along the line following the questions. Total score is number of introvert answers, the introvert section of the line being given by the scoring stencil.

*Norms:* Percentile norms for college students, men and women.

*Reliability:* .85 is given by Laird, N not stated.

### 2. Diagnostic Test for Introversion-Extroversion, by C. A. Neymann and K. D. Kohlstedt

*Date:* 1928.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Purpose:* To reveal tendencies toward introversion or extroversion.

*Designed for:* Adults.

*Contents:* Fifty statements to be answered by underlining Yes or No.

*Samples:*

Always be calm and collected. Yes No

Rewrite social letters. Yes No

Take an active part in all conversations going on around you. Yes No.

*Scoring:* The subject is requested to express his agreement or disagreement with the idea expressed in each statement by circling Yes or No. The number of questions answered in the extrovert direction are counted, and the number of introvert questions subtracted from this total. A minus score is introvert, therefore, and a plus score extrovert.

*Norms:* For 400 insane patients; 250 college students; 150 teachers, professional men, salesmen, etc. See Manual.

*Reliability:* Not reported.

### 3. Introversion-Extroversion in Terms of Interest, by E. S. Conklin

*Date:* 1923.

*Publisher:* University of Oregon, Eugene, Ore.

*Purpose:* To reveal introversion-extroversion through interests in contrasted activities.

*Designed for:* Adults.

*Contents:* Forty items to be rated on a nine-point scale.

*Scoring:* The score is the ratio of the sum of the reactions to extrovert items, to the sum of the reactions to introvert items.

*Norms:* None reported.

*Reliability:* .95 in a group of 352 college students.

*Reference:* CONKLIN, E. S., "The Determination of Normal Extrovert-Introvert Interest Differences," *Pedagogical Seminary and Journal Genetic Psychology*, 34:28-37, 1927.

#### 4. Introversion-Extroversion in Young Children, by L. R. Marston

*Date:* 1925.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Purpose:* To discover introvert tendencies in young children.

*Designed for:* Children two to six years old.

*Contents:* Twenty pairs of contrasting statements, the one representing the typical introverted, and the other the typical extroverted, attitude.

*Samples:*

- |  |   |
|--|---|
| ( ) Is self-conscious; easily embarrassed, timid or bashful.             | ( ) Is self-composed; seldom shows signs of embarrassment; perhaps is forward and "bold." |
| ( ) Rather insensitive and indifferent to others' opinions; independent. | ( ) Very sensitive and easily hurt; reacts strongly to praise or blame.                   |

*Scoring:* One member of each contrasting pair is to be marked ++ if it describes the child exactly, and + if fairly well. If neither statement describes the child both are marked -. Judgments are evaluated quantitatively as follows:

- 1 = ++ introvert trait  
 2 = + introvert trait  
 3 = - introvert trait-extrovert trait  
 4 = + extrovert trait  
 5 = ++ extrovert trait

The scale is scored cumulatively, scores ranging from 20 for extreme introversion, to 100 for extreme extroversion.

*Norms:* Average introversion-extroversion ratings and specimen profiles are given.

*Reliability:* .89 for six raters on comparable halves of scale, N = 26; .71 is average intercorrelation of three raters, N = 26.

*Reference:* MARSTON, L. R., "Emotions of Young Children," *Iowa Studies in Child Welfare*, 3:99, 1925.

#### 5. Personal Traits Rating Scale, by Edna Heidbreder

*Date:* 1927.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Purpose:* To measure (a) sense of inferiority, and (b) tendencies toward introversion.

*Designed for:* Adults.

*Contents:* Part I, consisting of 126 personality traits, is to be answered by S. Part II, consisting of fifty-four introversion-extroversion traits, is to be answered by S and by friends or associates.

*Scoring:* Part I: each item is designated as ++, +, 0, -, or --; from ++, indicating the presence of a trait to a marked degree, to --, indicating the presence of its opposite to a marked degree. The final score is the number of traits or symptoms of inferiority. Part II: each item marked + indicates introversion; - items indicate extroversion; and ? items indicate doubt. The final score is the algebraic sum of the + and - ratings.

*Norms:* For Part I, based upon 268 students in introductory courses in psychology; for Part II, based upon 200 students in general psychology.

*Reliability:* For Part I,  $r = .73$  (retest) 147 persons after an interval of six weeks; for Part II,  $r = .78$  (split-half)  $N = 200$ .

*Reference:* HEIDBREDER, EDNA, "Self Ratings and Preferences," *Journal Abnormal and Social Psychology*, 25:62-74, 1930-1931.

### III. ATTITUDE QUESTIONNAIRES

#### 1. A Test of Public Opinion (A Survey of Public Opinion on Some Religious and Economic Issues), by Goodwin Watson

*Date:* 1923.

*Publisher:* Bureau of Publications, Teachers College, Columbia University, New York.

*Purpose:* To show the extent and strength of an individual's prejudices, as exhibited in extreme opinions upon moral, religious and economic questions.

*Designed for:* High-school seniors and adults.

*Contents:* There are six parts or forms in this questionnaire. Form A is concerned with opinion upon social, religious and economic questions; Form B with the certainty of opinion in these same fields; Forms C, D, and E measure the extent to which an individual is willing to commit himself on moot questions; and Form F deals with the subject's willingness to generalize his opinions upon controversial issues.

*Scoring:* Extreme opinions, whether *pro* or *con*, are scored as evidencing prejudice. The gross total gives the general level of prejudice, and an analytic score is provided to indicate the direction of prejudice. A Manual gives directions for scoring.

*Norms:* Prejudice scores for various groups are given in the Manual.

*Reliability:* .96 (gross score) in a group of 161. The reliability coefficients of the analytic scores run from .60 to .88,  $N = 70$ .

*Reference:* WATSON, G. B., *The Measurement of Fair-mindedness*, Teachers College Publications, Columbia University, 176, 1925.

#### 2. Apperception Test, by Edith Burdick

*Date:* 1927.

*Publisher:* Association Press, 347 Madison Avenue, New York City.

*Purpose:* To measure the cultural background and social status of the child in the light of his attitudes and convictions.

*Designed for:* Children in the elementary grades.

*Contents:* Scale A contains eleven sections to be answered by the child as a group test. Scale B contains five sections to be answered by the child at home. These sections deal in general with knowledge of social usage, things liked and disliked, rôles of father and mother in the home, cultural knowledge, play activities, books read, and acts considered desirable.

*Scoring:* A scoring key was worked out from the combined ratings of competent judges.

*Norms:* Means and S.D.'s have been reported for a variety of groups.

*Reliability:* .62 to .77 in elementary grade groups. (N varied from 144 to 168.)

*Reference:* HARTSHORNE, H., AND MAY, M., *Studies in Deceit*, The Macmillan Company, New York, Book I, 205-212, 1928.

### 3. Experimental Study of Attitudes toward the Church, by L. L. Thurstone and E. J. Chave

*Date:* 1929.

*Publisher:* University of Chicago Press, Chicago, Illinois.

*Purpose:* To determine to what degree an individual is sympathetic or antagonistic toward the modern church.

*Designed for:* Adult groups.

*Contents:* Forty-five statements ranging from those very favorable to the church to those very unfavorable.

*Scoring:* S is to check those statements with which he agrees. Final score is determined from the average of the scale values of the opinions expressed.

*Norms:* Results for several groups are given: freshmen, sophomores, juniors and seniors, graduate students, divinity students, churchgoers, non-churchgoers, Jews, Protestants, Catholics, men and women.

*Reliability:* .92 (split-half), N = 200 freshmen.

*Reference:* THURSTONE, L. L., AND CHAVE, E. J., *The Measurement of Attitude*, University of Chicago Press, Chicago, 1929.

### 4. Test of International Attitudes, by G. B. Newmann, D. H. Kulp and Helen Davidson

*Date:* 1926.

*Publisher:* Bureau of Publications, Teachers College, Columbia University, New York.

*Purpose:* To discover international and interracial attitudes.

*Designed for:* Young people and adults.

*Contents:* Four sections, A, B, C and D, dealing with attitudes toward international questions, beliefs and convictions regarding other races and nations. In Section A each item is preceded by a + ? -; items in other sections are preceded by R+ R ? W W-.

*Scoring:* Agreement or disagreement is expressed by circling the appropriate symbol. Final score is total number of points obtained on all items (each answer has a different weight which is given by the key), divided by 108, the number of items in the whole test.

*Norms* Given in the Manual for various high-school groups.

*Reliability:* .77 to .98 (split-half) within a group of 346 high-school seniors.

5. Test for Social Attitudes and Interests, by Hornell Hart

*Date:* 1923.

*Publisher:* University of Iowa, Iowa City, Iowa.

*Purpose:* To indicate predominant attitudes, *i.e.*, likes and dislikes toward a variety of activities, and toward a number of social and economic situations.

*Designed for:* Adults and adolescents over twelve years.

*Contents:* Chart 1 contains practice lists; Chart 2, four lists of activities, social, emotional, religious, economic, for which preference or lack of preference is to be expressed. Chart 3 contains two lists of things to read or study and two lists of economic and social reforms upon which the subject is to express an opinion. Each item is followed by a + and -.

*Scoring:* S is to circle the + if the item is liked, and the - if it is disliked. The liberal or socialized responses were determined from a group of "leaders of social progress," with which S's responses may be compared.

*Norms:* The responses of social leaders and other groups are supplied by the author.

*Reliability:* Not reported.

*Reference:* HART, H., *A Test of Social Attitudes and Interests*, University of Iowa Studies in Child Welfare, II, 4, 1923.

#### IV. INTEREST QUESTIONNAIRES

1. Interest Questionnaire for High School Students, by O. K. Garretson and P. M. Symonds

*Date:* 1931.

*Publisher:* Bureau of Publications, Teachers College, Columbia University, New York.

*Purpose:* "To supply a valid and reliable measure of the inclination of pupils entering high school toward the academic, commercial, and technical curricula, through a sampling of their preferences over a wide range of items."

*Designed for:* High-school boys.

*Contents:* Eight tests dealing with occupations, sports, student school activities, school subjects, things desired, magazines read, *etc.* S expresses preference by circling L, I, D. Samples:

Test 1—Barber	L	I	D
Test 3—Outdoor work	L	I	D
Test 5—Member of glee club	L	I	D
Test 8—Detective stories	L	I	D

*Scoring:* Test is scored separately for academic, commercial and technical interests. Final score is sum of credits given items.

*Norms:* Decile scores for academic, commercial and technical interests based on questionnaires from approximately 800 ninth-grade boys.

*Reliability:* .86 (split-half) for academic preference; .93 (split-half) for commercial preference; .95 (split-half) technical preference. Each reliability coefficient is based upon twenty-five cases.

## 2. Occupational Interest Blank for Women, by Grace E. Manson

*Date:* 1931.

*Publisher:* School of Business Administration, University of Michigan, Ann Arbor, Michigan.

*Purpose:* To determine whether a given woman's interests serve to identify her with a well-defined occupational group.

*Designed for:* Women.

*Contents:* This blank contains 160 occupations, open to women, arranged alphabetically. Each occupation is followed by five symbols L! L ? D D! which denote different degrees of "liking" and "disliking."

*Samples:*

Auditor	L!	L	?	D	D!
Interior decorator	L!	L	?	D	D!
Singer	L!	L	?	D	D!

*Scoring:* Stencils in which each choice is given a + or - numerical weighting are employed as in the Strong Vocational Interest Blank. There is a different stencil for each occupation. If a blank is scored for "trained nurse" interests, the total score classifies the subject as having the interests of a nurse, as not having the interests of a nurse, or as doubtful.

*Norms:* Scales are available for ten occupational interests, e.g., high-school teacher, private secretary, stenographer.

*Reliability:* .89 (split-half) the average of ten reliability coefficients, each based upon a sample of fifty drawn from the given occupation.

*Reference:* MANSON, GRACE E., *Occupational Interests and Personality Requirements of Women in Business and the Professions*, Michigan Business Studies, 3:281-409, 1930.

## 3. Specific Interest Inventory, Form W, by Frances J. Stewart and Paul Brainard

*Date:* 1932.

*Publisher:* Psychological Corporation, 522 Fifth Avenue, New York City.

*Purpose:* To aid in the discovery and analysis of interests, with a view toward vocational guidance.

*Designed for:* Various groups, Forms B, G, M and W, for boys, girls, men and women, respectively.

*Contents:* Twenty groups of questions plus an inventory of likes and dislikes. Each group contains five questions covering different phases of a given mode of expression. These are scored under Dislike, Neutral (N), Like. For example, under music are included instrumental and vocal performance, analysis, composition, and appreciation of music.

*Samples:*

Manual (expression). How do you like—

	Dislike		N	Like	
41. To wrap parcels, cover or repair books, making neat folds?	1	2	3	4	5
42. To do fine needlework, make or repair clothes?	1	2	3	4	5
43. To trace patterns or maps; to draw with careful measurement?	1	2	3	4	5
44. To cut out patterns; to shape objects carefully with knife, scissors, or scroll saw?	1	2	3	4	5
45. To do weaving, basketry, wood-work, varnishing, painting, etc.	1	2	3	4	5

*Scoring:* Final score is sum of all numbers circled. Subtotals, found by totaling numbers circled under each of the different headings, are arranged in order of size in order to provide an analysis of an individual's interests. Other methods of analyzing the scores from the blank into vocational patterns are suggested in the Manual.

*Norms:* No average scores, or norms, are given for different groups. An analysis of scores, however, indicates the order and range of interests. From this, tentative educational and vocational possibilities are suggested.

*Reliability:* Shifting of interests was studied by giving the Inventory to sixty ninth-grade children in November and again in May. The self-consistency of the twenty subgroups, as shown by their  $r$ 's, ranged from .13 to .94, with an average at .68.

#### 4. Vocational Interest Blank, by E. K. Strong

*Date:* 1927.

*Publisher:* Stanford University Press, Stanford University, California.

*Purpose:* To discover whether an individual's interests are closely identified with those of some well-defined occupational group, e.g., lawyers, chemists, physicians.

*Designed for:* Adults.

*Contents:* This blank contains eight parts. Parts I, II, III, IV and V deal respectively with occupations, amusements, school subjects, activities and peculiarities of people. Parts VI, VII and VIII deal with prefer-



ences for various activities, careers as typified by well-known men, comparisons of interests, abilities and characteristics. S answers Parts I–V by circling the L (like), I (indifferent) or D (dislike) after each item. In Part VI, the first three and last three choices are indicated by check marks; and in VII and VIII answers are indicated in the appropriate column. Samples are:

Part	I—Astronomer	L	I	D
	Farmer	L	I	D
Part	II—Golf	L	I	D
	<i>Atlantic Monthly</i>	L	I	D
Part	III—Algebra	L	I	D
Part	IV—Making a radio set	L	I	D
Part	V—Witty people	L	I	D
Part	VIII—Get rattled easily	Yes	?	No

*Scoring:* Blank is scored separately for each occupation. Final score for a given occupation is algebraic sum of weights assigned to L, I, D. Detailed directions are given in Manual.

*Norms:* When scored for a given occupation, say farmer, S's total score enables him to be assigned a rating of A (having interests of farmer), B (not sure), C (does not have interests of a farmer).

*Reliability:* .80, average of twelve reliability coefficients obtained from eighty-three lawyers, fifty architects, forty-five C.P.A.'s, 100 college students, scored for lawyer, architect, C.P.A. scales.

## THE VALIDITY OF THE QUESTIONNAIRE

### 1. General Principles

We have previously shown (p. 124) how Thurstone and other investigators used the criterion of "internal consistency" in the selection of the items to be included in the final forms of their questionnaires. This technique is clearly a method of validation as well as a method of selection. It loads the questionnaire with items known to be discriminatory; and furthermore when items exhibit self-consistency, *i.e.*, show close agreement among themselves, we know that the questionnaire is, at least, tapping related aspects of behavior.

Internal consistency is a necessary preliminary method in the standardization of a questionnaire. Once the questionnaire has been completed, however, its validity, as a finished instrument, may be considered from two points of view. The first is concerned with the veracity of the subject's report; the second with the agreement of the questionnaire with objective measures of what it is attempting to reveal. We may consider these in order.

When one has gathered together a list of questions agreed upon by experts as furnishing a comprehensive sampling of the field studied, the validity of a subject's report will obviously depend upon his veracity. If the conditions of the experiment are good, and the motivation such as to insure reasonable honesty, an individual's report must be taken at face value. There is, in a sense, no more valid report than a person's own statements concerning himself. Suppose, for example, that a man reports that he is timid, and is bothered by fluttering of the heart; or that he is a fundamentalist in religion; or is interested in medicine rather than in law. The question of whether this individual actually possesses the symptoms, or the attitudes and interests which he describes, is of no more importance than the fact that he *believes* that he possesses them. That neurotic symptoms do not always exhibit themselves in behavior is no argument against their reality. Feelings of social inferiority, the dislike of people and crowds, *etc.*, are often reported by persons who, to casual observation, seem perfectly at ease in social situations. Again, a change in motivation may radically affect a person's report. Hollingworth (31) found a drop of from fifteen to twenty points in the number of unfavorable answers to the P.D. Sheet given by pre- and post-armistice soldier groups. There is no more reason to believe that these men lied than there is to believe that both reports were valid (*i.e.*, truthful), since the second report was influenced by conditions radically different from those present in the first.

The usual method of determining the validity of a completed questionnaire is to check its data against the judgment of experienced observers; or to find the correlation of the questionnaire with other determinations of the traits which the questionnaire is trying to evaluate. Franz (17), commenting on the Woodworth P.D. Sheet, and basing his opinion upon clinical findings, states that "probably any individual who answers twenty of the questions wrongly should be suspected of instability. If the number of wrong answers is greater than thirty, grave suspicion of abnormality is warranted." Laird (37) says of his Personal Inventory B2 that ". . . it has found a large number for whom there has been a distinct need for orthosis, and who would otherwise probably never have been noticed until something serious had occurred." Woodworth reports that normals had a

median score of 10 unfavorable answers on the P.D. Sheet, while psychoneurotics gave from 30 to 40 unfavorable answers.

## 2. Personality and Adjustment Questionnaires

Mathews (39) had four competent judges rate each of thirty-five girls in a Protectorsy for "nervous instability," using a four-point scale. The correlation between these composite ratings and the Mathews revision of the P.D. Sheet was .52; between composite ratings and a second giving of the P.D. Sheet .66 ( $N = 28$  girls). Cady (9) has reported correlations of .41, .42, and .36 between his revision of the P.D. Sheet and teachers' estimates of incorrigibility. Cady's subjects were a group of 150 incorrigible boys confined in institutions. Terman (62), using the Cady revision with some additions, has studied the emotional stability of gifted children. These children, age for age, far surpassed the normal control group, both in absence of nervous symptoms and in social adjustment. Slawson (54) arranged the seventy questions in the Mathews revision of the Woodworth P.D. Sheet in order in accordance with the number of unfavorable answers given by the boys in three institutions for the delinquent. The average correlation between the ranks for the three groups was .91. This high agreement in the incidence of reported symptoms is to be contrasted with the correlation of .55 between the orders of the questions for normal children. Such results as these indicate that the Woodworth P.D. Sheet and its revisions sift out, with fair accuracy, persons handicapped or liable to be handicapped by poor social and personal adjustments.

The Thurstones (67) report results secured with the Personality Schedule which have a distinct bearing upon its practical usefulness. The more neurotic students, as judged by the number of unfavorable answers, do better academic work than those presumably better adjusted. This result has been substantiated by Laird and by Fleming. On the average, women score higher than men on the schedule, non-fraternity men higher than fraternity men, Jewish students higher than Gentile students. Bridges (8), who administered the P.D. Sheet to 136 college men and thirty-two women, reports women to be less stable emotionally than men. Students also prove to be less stable than the general population. The typical student neurosis, according to Bridges, is characterized by anxiety, irritability, worry and disturbed sleep.

The correlation of the Thurstone Schedule and the Bernreuter Inventory (4) scored for neurotic traits was .94 in a group of seventy students in a class in elementary psychology. Bernreuter reports correlations of .76 ( $N=70$ ) and .69 ( $N=44$ ) between his Inventory when scored for introversion and the Colgate Introversion Test C2. Both of these results were based upon data obtained from college students. The Allports (2) have attempted to validate their A-S Study against ratings for ascendance-submission. The correlation between self ratings, in their standardization group of 400 men, and scores on the A-S test was .63; between ratings by associates and test scores .46. These correlations are spuriously high because the same group that was used in determining the scale values was also employed in validating the test. In later studies,  $r$ 's of .29, .30, and .33 were obtained between ratings for ascendance-submission and scores on the A-S Study in groups of forty-two men, fifty-one women, and twenty-one men and women, respectively. These  $r$ 's do not indicate a high degree of validity for the test, although the probable unreliability of the ratings precludes a definite conclusion. The authors comment that "the ultimate validity of the study will in all probability be established only in terms of its practical success in vocational guidance, clinical and personnel work and other forms of personality study."

The Pressey X-O Test contains so many diverse elements that a "total affectivity" score has little value or meaning. The interpretation of an "idiosyncrasy" score is also doubtful, since "modal choices" of words in the various lists fluctuate greatly depending upon the group used in the standardization. Pressey (46) has recognized these facts and has emphasized the need for "differential scores" based upon an analysis of the items separating two contrasted groups. Chambers (10), working with 200 college students, devised a differential score for the X-O test by selecting those marked-out words which separated the high 25 per cent. ("good" students) in the grade distribution from the low 25 per cent. ("poor" students). A score of +1 was given a student for each word crossed out which was characteristic of the high 25 per cent., and a score of -1 for each word crossed out which was characteristic of the low 25 per cent. The final score was the algebraic sum of the plusses and minuses. This "net differential score" correlated .54 with grades, while the correlation between grades and intelligence tests

in the same group was only .33. It should be noted that the  $r$  of .54 is partly spurious as the differential scores were derived from the grade distribution, and hence are directly dependent upon them. However, in another group of fifty-seven students, Chambers obtained an  $r$  of .46 between the same two measures which indicates considerable promise for the method. In another study, Chambers (11) has applied the method of differential scores with considerable success to the study of emotional maturity in children from Grades 6 to 12.

### 3. Introversion-Extroversion

A correlation of  $-.22$  between the introversion and extroversion scores of the Colgate Mental Hygiene Test C1 has been reported by Hoitsma (30) in a group of 288 men who took the test twice, and a correlation of  $-.45$  between the same two tests in a group of 268 women. These  $r$ 's indicate considerable opposition in the two attitudes, but not enough to suggest entirely opposed types. In the same study, the correlation between introversion and scholarship in a group of 218 men was .35, and the correlation between introversion and the Colgate Mental Hygiene Test B1 was .49. The indication is that the better students tend to be introverted and also more susceptible to nervous symptoms.

### 4. Attitudes

G. B. Watson (71) has tested the validity of his fair-mindedness test in various ways. The homogeneity of the test as a whole was investigated by finding the correlation of each of the six parts with the total score on the test. These correlations range from .53 to .94 ( $N=40$ ) with an average of .68, indicating a fairly substantial community of attitude within the test as a whole. Case studies were also made of the scores of individuals and groups considered by their associates to be especially fair-minded or unprejudiced, and of other individuals and groups believed to be prejudiced along definite lines. In these comparisons the "prejudiced" individuals returned scores from 35 to 65 per cent. higher than the "unprejudiced." These results indicate a considerable degree of validity for this test.

E. K. Strong (57) has validated his Vocational Interest Blank against several criteria. The construction of scoring keys for each occupation separately is, of course, a direct means of validation,

since scoring weights (for lawyer, engineer, *etc.*) are based directly upon the expressed interests of the particular groups concerned. The validity of the test was studied further by comparing the dominant interests registered by men taking the test and their later vocational choices. The Vocational Interest Blank, for example, was administered to 284 Stanford University students at the beginning of their senior year, and nine months later upon graduation a report was secured from those who had made a definite choice of occupation. Strong found that 71 per cent. of this group had selected occupations in which they rated highest or second highest in interest. Six months after graduation reports were secured from 156 of these students concerning the occupations which they had entered or were planning to enter. Nearly half of these men were entering the occupation in which their interest scores were highest, and 77 per cent. were entering the occupation in which their interest scores were first, second or third highest. If men tend in general to enter those occupations in which they are interested, these results indicate a high degree of validity for the Vocational Interest Blank. Still another common-sense method of validation was used by Strong (58). This was to compare the interests of vocations which seemed to appeal to the same—or to very different—groups. The results here are fairly consistent with common experience. The interests of the engineer, for instance, were found to be closely related to the interests of the chemist or farmer, and remotely related to the interests of the minister or the advertising man.

## 5. Interests

Garretson (24) has obtained results which indicate that his interest questionnaire will predict quite accurately the subjects which a boy will select in high school. Using the bi-serial correlation method (35), Garretson found a high correlation between specific interests and choice of school studies involving these interests. Thus, the correlation between interest expressed in commercial subjects and the choice of this curriculum was .73; between technical interests and the choice of technical subjects .87; and between academic interests and the choice of academic subjects .56. No correlation was found by Garretson between commercial, technical or academic interests and general intelligence.

Wyman's Free Association Tests of Interest (72) is an ingenious

attempt to measure interests objectively through the free association technique. This test consists of two sets of sixty stimulus words each, the responses or associations to which can be scored separately for intellectual, social or activity interests. According to Wyman, those responses which exhibit, predominantly, interest in "knowing" are adjudged intellectual; those showing predominantly interest in persons, social; and those showing interest in "doing," activity. In constructing scoring keys for her test, Wyman weighted each response for intellectual, social or activity interests, according to the frequency with which each response was given by children selected by their teachers as having primarily intellectual, social or activity interests. The validity of the test was investigated by finding the correlations of interest scores with teachers' ratings for intellectual, social and activity interests in groups of children numbering twenty-one to seventy-eight. The average corrected correlations were .65, .50 and .31, respectively, between intellectual, social and activity interests ratings. Besides indicating a fair degree of validity, these results suggest that either the interest test or the teachers' ratings (probably the latter) select most effectively children with intellectual, *i.e.*, academic or school, interests.

The correlations of intellectual, social and activity interests and school achievement as measured by the Stanford Achievement Test were .63, .50 and .40, respectively, in a group of eighty-one sixth- and seventh-grade pupils. When the variability due to differences in intelligence scores was held constant, these correlations become .49, .18 and .03, showing the relatively much greater importance of intellectual interest in school work.

Nearly all investigators agree in finding little correlation between personality measures and general intelligence test scores. Hoitsma (30) reports zero or close to zero correlation between the Colgate B1 (early form of B2) Test and general intelligence and scholarship. Flemming (16) obtained  $r$ 's between the Thorndike Intelligence Test and the Woodworth P.D. Sheet, the Colgate B2 and the Pressey X-O of .01, —.004 and .07, respectively, in a group of approximately 300 freshmen. Conklin (12) obtained a correlation between general intelligence and introversion of .05 in a group of 159 college students. Thurstone (68) obtained a correlation of .03 between his Personality Schedule and general intelligence in a group of 694 college students.

## OBJECTIVE TESTS OF PERSONALITY AND CHARACTER

In this section several representative objective tests, designed to measure personality and character, will be described, and a summary given of the extensive battery of tests developed by the Character Educational Inquiry (C.E.I.) under the direction of Harts-horne and May and their collaborators. These latter tests furnish perhaps the best illustration of the construction and use of objective tests in the study of character, ethical and moral attitudes. Other earlier and ingenious objective tests in the same field have been devised by Raubenheimer (47) and Voelker (70).

Objective personality tests differ from questionnaires and rating scales in that they set before the subject tasks which are scored in terms of *amount done*, or *time taken to complete*. In some instances, to be sure, the classification of a measure as a questionnaire or as a test is largely a matter of convenience. The terms are often used interchangeably, and there is in many cases little real distinction between the two types of measurement. The following paragraphs will present a summary of several types of personality tests. Following this, an evaluation of the validity and use of these measures will be attempted.

## I. PERSONALITY TESTS

## 1. Ethical Discrimination Test, by S. C. Kohs

*Date:* 1922.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Purpose:* To measure knowledge of ethical and moral principles.

*Designed for:* Children and adolescents; may be used with adults.

*Contents:* This test contains six "exercises" or sub-tests, as follows:

Exercise 1. Social relations.

Exercise 2. Moral judgment.

Exercise 3. Proverbs.

Exercise 4. Definition of moral terms.

Exercise 5. Offense evaluation.

Exercise 6. Moral problems.

*Scoring:* The score on each test is number right divided by number attempted, except in exercise 4, where the score is number right over 45. Final score is sum of separate scores.

*Norms:* Average scores based upon approximately 100 cases are given in the Manual for ages six to seventeen. Also a tentative classification of individuals according to score, into moral imbecile, moral deficient, inadequate, sub-average and so forth, is provided.



*Reliability:* No data reported.

*Reference:* KOHS, S. C., "Ethical Discrimination Test," *Journal Delinquency*, 7:1-15, 1922.

**2. Group Will-Temperament Tests, by June Downey (an individual form of this test is also published)**

*Date:* 1922.

*Publisher:* World Book Company, Yonkers, New York.

*Purpose:* To evaluate certain temperamental traits in individuals through a study of simple motor reactions, usually handwriting.

*Designed for:* Children and adults.

*Contents:* A series of twelve tests, arranged into three groups, or patterns, of four tests each. The separate tests require the subject to mark that one of a series of pairs of contrasted adjectives which best describes himself, to write as rapidly or as slowly as possible, to practice copying a model, to write with eyes closed, to write under forms of distraction, etc.

*Scoring:* Detailed directions for scoring (with illustrations) are given in the Manual. By means of tables all scores may be expressed upon a scale which runs from 1 to 10. A profile or graphic representation of his scores gives the individual's rank in the various traits. According to the author, high scores in the first four tests classify a person as of the "explosive, hair-trigger" type, high scores in the second four tests as willful and aggressive; and high scores in the last four tests as slow, accurate and tenacious.

*Norms:* A subject's relative position in the various traits is given on his profile.

*Reliability:* .60 for the whole scale, individual reliability coefficients ranging from .30 to .80, roughly. These reliability correlations are averages based upon retests made after a one-day interval. The groups tested were 149 normal school women, forty-two high-school boys, and thirty-seven high-school girls (15). Ruch and Del Manzo (50) give .38 as the reliability of the whole test in a group of 146 high-school students.

**3. Social Intelligence Tests (Revised Form) by F. A. Moss, T. Hunt and K. T. Omwake**

*Date:* 1927.

*Publisher:* Center for Psychological Service, Washington, D. C.

*Purpose:* To measure a person's "social intelligence, defined as ability to deal with people."

*Designed for:* Adults and adolescents.

*Contents:* Five tests described as follows:

- (1) Judgment in social situations.
- (2) Recognition of the mental state of the speaker.
- (3) Observation of human behavior.
- (4) Memory for names and faces.
- (5) Sense of humor.

*Scoring:* The final score is the sum of the points earned on the five separate tests. Directions for scoring accompany the test.

*Norms:* Percentile norms for college freshmen and upper classmen are given, based on approximately 2,000 cases.

*Reliability:* .89 (retest) for 100 sophomores; .88 (split-half) for 129 college students. These reliability coefficients are for the 1925 edition of the test.

*Reference:* HUNT, T., "The Measurement of Social Intelligence," *Journal Applied Psychology*, 12:317-334, 1928.

#### 4. Suggestibility Test for Children, by Margaret Otis

*Date:* 1924.

*Publisher:* The C. H. Stoelting Company, Chicago, Illinois.

*Purpose:* To study "ability to resist suggestion" in children, by means of a group paper-and-pencil test.

*Designed for:* Elementary and high-school children, primarily.

*Contents:* Two forms, A and B, each form containing forty simple tasks. Twenty questions in each form are worded so as to suggest a certain response, the suggestion being sometimes direct, and sometimes indirect. These test items are alternated with twenty simple directions items, the idea being to conceal the purpose of the test. The test is labeled "Directions Test."

*Scoring:* Each successful resistance, i.e., refusal to accept a suggestion or implied response, is scored 5; a perfect score is 20 times 5, or 100. Scoring keys and directions are provided.

*Norms:* Distributions of suggestibility scores by grade, by M.A. and by C.A. are provided.

*Reliability:* .58 (Form A against Form B),  $N = 831$  elementary school children.

*Reference:* OTIS, MARGARET, "A Study of Suggestibility of Children," *Archives Psychology*, 70, 1924.

## II. CHARACTER EDUCATION INQUIRY TESTS

This extensive battery of tests may be conveniently classified under several sub-heads in accordance with the general character of the behavior studied. The first group of tests, those designed to measure honesty and trustworthiness, will be presented in some detail; the other batteries will be described in a more cursory fashion. For a detailed description of the individual tests, the reader is referred to the references given below.

### 1. Tests of Honesty and Trustworthiness, by Hugh Hartshorne and Mark May

*Date:* 1928.

*Publisher:* Association Press, 347 Madison Avenue, New York City.

*Purpose:* To devise tests for discovering (a) the incidence and degree

of cheating, stealing and lying in various groups of children; and (b) the relation of these forms of behavior to such variables as age, sex, intelligence, religious instruction and various personality measures.

*Contents:* The tests of honesty may be classified into nine groups, as follows:

(1) I.E.R. (Institute of Educational Research) Tests.

There are four tests in this group: Arithmetic; Sentence Completion; Information; and Word Knowledge. Each of these tests has two forms which were experimentally equated in difficulty. Both forms of each test were given to experimental groups, the first form under conditions which did not permit of cheating, the second form under conditions in which cheating could easily take place. The cheating situation was arranged by allowing each pupil to score his own paper by means of an answer key. In order to provide comparisons, the same two forms of each test were then given to various control groups, under conditions wherein cheating was impossible. This procedure allowed E to estimate statistically how much variation in score from the first to the second testing might occur under honest conditions. If the difference between the scores in the experimental groups from the first to the second testing exceeded the normal variation to be expected (as determined by results from the control groups), cheating presumably occurred. This procedure is known as the "*double testing technique*."

(2) Speed Tests.

There were six tests in this group, classified as follows:

- (1) Addition of one- and two-digit combinations.
- (2) Number-checking tests.
- (3) "A" cancellation test.
- (4) Digit-symbol substitution test.
- (5) Dot in square test.
- (6) Digit-cancellation test.

These tests were given twice under honest conditions, one minute being allowed for each trial of each test. On the third trial, each child was allowed to score his own paper. Those inclined to be dishonest could easily add on more items to their papers, thereby increasing their scores unfairly. The probability of an unfair increase was determined by comparing the tests performed under honest and under cheating conditions.

(3) Coördination Tests.

The coördination tests consisted of three tracing tests, a squares and circles test, and a maze test (for detailed description see [26]). The subjects were instructed to perform specified tasks with their eyes closed. If a child peeped, in this way improving his score beyond the limits of probable achievement (these limits were determined by control groups), he was scored as hav-

ing cheated. This method is called "*the improbable achievement technique*."

(4) Puzzle Tests.

There were three puzzle tests. These devices, while appearing simple, were really quite difficult. A subject was scored as having cheated if he faked the solution of a puzzle. This is again the method of improbable achievement.

(5) Lying Tests.

Two tests were constructed to reveal deception by lying. The first was given a week or more after the I.E.R. tests and consisted of a number of questions, some of which asked directly if the subject had cheated on the I.E.R. tests. The truth or falsity of the child's reply could be checked against his actual record. This test was called "lying to escape disapproval." The second test of lying consisted of two forms of thirty-six questions each. These questions covered a variety of specific acts, which, although they have widespread social approval, are rarely actually performed. Samples are:

12. Do you always smile when things go wrong? Yes No

19. Do you always obey your parents cheerfully and promptly? Yes No

27. Do you go to church and Sunday School every Sunday? Yes No

32. Do you usually correct other children when you hear them using bad language? Yes No

Children who made scores in the conventionally correct direction, beyond certain standards set by an adult group of graduate students who answered the questions in the way which they felt truly represented their childhood, were adjudged to have lied to gain approval.

(6) Home Work Tests.

One form of the I.E.R. word-knowledge test was handed to each child with the request that he fill it out at home and bring it in next day. Children were told twice not to get any help, either from the dictionary or from another person. Scores on this home test were compared with the scores made on the equivalent form of the test given in school under honest conditions. Cheating was judged to have occurred when the difference between these two scores was greater than the normal variation to be expected on a retest.

(7) Athletic Contests.

In the Athletic Contests four measures of physical ability were employed: the dynamometer; the spirometer; the chinping or pull-up test; and the standing broad jump. Motivation was secured by offering an attractive badge designating a boy, say, as

"grip-champion," or a ribbon designating a girl as "pull-up champion," of a particular grade. The factor of social inhibition, that is, unwillingness to compete against a group, was eliminated to a large degree by making the tests entirely individual. Each child was allowed three trials of a test, the best one out of the three being noted mentally by the examiner without the child's knowledge. The children were then encouraged to make five additional trials. These were done without any supervision, each child recording his own results, and reporting them later to the examiner. Since the practice effect in these physical tests is almost nil, and the fatigue effect considerable, it is highly improbable that a child will report a better record as a result of additional trials. Hence, any reported gain over and above an established normal variation was considered evidence of deception.

(8) Parties Tests.

In the Parties Tests various parlor games, popular with children, *e.g.*, pinning the tail on a donkey, bean relay race, *etc.*, were utilized. Cheating can easily occur in these games. If the child took unfair advantage, or reported results falsely, he was scored as having deceived.

(9) Money-taking Tests.

The purpose of these three tests was to see if a child would steal a small amount of money if given the opportunity without any apparent danger of detection. These tests, which are quite ingenious, are described in detail in (26).

*Norms:* The number and percentage of children who were deceptive are given for all tests and for various groups from the fourth grade through the high school in HARTSHORNE, H., AND MAY, M., *Studies in Deceit*, Book 2, chapter 4, 1928.

*Reliabilities:* Reliability coefficients are reported for all tests except the Parties and the Money-taking Tests. Results may be summarized (26) as follows:

	<i>r</i>
(1) Reliability of I.E.R. tests (Arith., Sen. Comp., Inform.)	.86
(2) Reliability of Home Test (Word knowledge)	.24
(3) Reliability of Speed Battery (Six tests)	.83
(4) Reliability, physical ability tests (battery of four)	.77
(5) Reliability, three coördination tests	.72
(6) Reliability, three puzzle tests	.75
(7) Reliability, lying tests	.84

## 2. Other C.E.I. Tests

### (1) Tests of Cooperation (27).

This group consists of five tests designed to measure coöperation and helpfulness. In a general way the plan was to measure the relative efficiency of children when working on a group project and when working for personal benefit.

### (2) Tests of Inhibition (27).

These tests, six in number, were designed to measure a person's ability to exercise self-control while working at a monotonous task. Self-control, or inhibition, was measured by the subject's ability to resist the distraction of interesting pictures and reading material, and to postpone the desire to manipulate interesting play-things and puzzles until the assigned task was accomplished.

### (3) Tests of Persistence (27).

These tests were attempts to measure the willingness to persist in work of an uninteresting sort when the subject was permitted to stop whenever he liked. The tasks required that the child read a story in which words were presented without spacing, or to keep on trying to solve difficult puzzles. All of the tasks demanded a high degree of attention and effort. The child was motivated in two ways to persist in his efforts: first, to improve his own score, secondly, to improve the record of his class.

### (4) Tests of Moral Knowledge and Attitudes (28).

This group of twelve tests may be classified into four groups according to the nature of the task set, as follows:

- (a) Comprehension. Situations presenting problems were described, each suggesting four solutions. The child was required to indicate what in his opinion were the most sensible, the most helpful and the most useful alternative solutions.
- (b) Information. Tests of knowledge of words having moral significance; knowledge of the differences between lying, stealing and cheating; foresight of consequences; cause and effect relations.
- (c) Opinions. These tests had to do with duties, principles, and the importance of consequences.
- (d) Attitudes. These tests were concerned with attitudes toward misconduct, and generalizations about moral conduct.

## THE VALIDITY AND USE OF PERSONALITY AND CHARACTER TESTS

One of the greatest obstacles which the investigator encounters in trying to validate an objective personality test is the difficulty of securing independent measures of the traits to be studied. The Downey Will-Temperament Test furnishes a good example of what is meant. Several attempts have been made to validate the will-

temperament tests by computing the correlation between test scores and ratings of the same qualities ostensibly measured by these tests. Meier (41) calculated the correlation between scores on the twelve separate traits measured by the Downey Will-Temperament Test (individual form) made by 106 high-school students, and ratings for the same traits made upon individuals in this group by at least three judges, *e.g.*, teachers, parents or friends. The amount of agreement between ratings and scores was extremely low, the average correlation being .12. The highest  $r$ , *viz.*, .24, was between ratings and scores for motor inhibition. Ruch and Del Manzo (50) computed  $r$ 's between scores made on the twelve traits presumably measured by the Will-Temperament Test and ratings by three to four teachers upon a specially selected group of twenty-eight high-school students. These subjects were selected from a larger group of 146, and represent, in equal proportions, the two extremes of the group when rated for "will power." If there is any relationship between will-temperament scores and ratings it would seem certain to appear in a group selected to show marked contrasts in "will." Actually, however, the average  $r$  between ratings and scores was .15, the separate  $r$ 's ranging from .53 to —.29.

The lack of correlation between ratings and scores on the Will-Temperament Test results in part, at least, from the unreliability of the criterion ratings. It is very difficult to rate children on the traits presumably measured by the will-temperament tests. Such characteristics as flexibility, motor impulsion, interest in detail, and volitional perseveration, are difficult to define as well as to measure, and there is probably little understanding by the individual rater of what such general terms mean. Again, the function utilized by the will-temperament tests, *i.e.*, handwriting, is extremely narrow, so that test scores and ratings are probably estimates of very different abilities. The highest  $r$ 's between ratings and scores (49) are for those tests involving speed of movement—the most objective of the will-temperament tests.

Other attempts at validating the will-temperament tests have been through the intercorrelations of the tests themselves, and through a study of individual profiles. High intercorrelations indicate that the members of a test battery are presumably measuring some common trait or traits. Again, if a man's profile identifies him as belonging to the slow, accurate, tenacious "type" and his friends agree upon

this description, the test is, for this case at least, valid. Unfortunately, the intercorrelations of the Downey Will-Temperament Test have not been high. Ruch and Del Manzo (50) report the average intercorrelation of the four tests designed to measure the "explosive, hair-trigger type" to be .22; of the four tests to measure the "willful, aggressive type," —.06; of the four tests designed to measure the "slow, tenacious type," .07. The subjects were 146 high-school students. Uhrbrock (69) reports many correlations between various groupings of the Will-Temperament Test. All of these are low, however, none being above .60, and several being zero or negative. Downey (14) has employed the case study method of validation, in which judges were requested to identify the profiles of persons known to them. Several rather striking successes were recorded, but the percentage of successes is by no means high, and the validity of the Downey Will-Temperament Test is hardly established by this method. On the whole, we must conclude that the Downey Will-Temperament Test is not, at present, a satisfactory measure of any clearly defined personality traits. Much further experimental work with this test is necessary before it can be used with assurance.

The George Washington University Social Intelligence Test represents an attempt to use the general intelligence test technique to measure a person's knowledge of social amenities, his interests in people, and his insight into social situations. The test suffers from the fact that it is of the paper-and-pencil type, and that the situations which it sets up are hypothetical instead of actual. It is a matter of observation that social behavior is to a large extent a matter of habit; a person may know the correct thing to do, and still not act in accordance with his expressed or written decisions. Correlations computed with general intelligence tests indicate that the George Washington Social Intelligence Test is to a fair degree a measure of abstract intelligence. Hunt (34) reports a correlation of .54 between the George Washington University Mental Alertness Test and the Social Intelligence Examination, in a group of 243 college freshmen. Garrett and Kellogg (23) obtained a correlation of .42 between the Thorndike Intelligence Test and the Social Intelligence Test in a group of 118 freshmen. Hunt reports correlations of .22 between the O'Rourke Mechanical Aptitudes Test and Social Intelligence in a group of 130 high-school pupils, and an  $r$  of .11 between the



McQuarrie Mechanical Ingenuity Test and Social Intelligence, in a group of 176 college students.

Hunt has attempted to validate the social intelligence test by studying the scores for social intelligence in relation to the student's extracurricular activities. The assumption is that a large number of outside activities will indicate a high degree of social intelligence. For a group of 262 freshmen, taking full-time work, the median social intelligence test score in relation to activities was:

	Average Social Intelligence
4 or more activities .. . . .	116
3 activities . . . . .	112
2 activities . . . . .	110
1 activity . . . . .	105
0 activities .. . . .	99

No measure is given of the reliability of the differences between these groups, but the trend indicates that the more "social-minded" students (as measured by the test) tend on the whole to engage in more outside activities. It seems clear that the George Washington Social Intelligence Test measures aspects of behavior which are neither strictly abstract nor mechanical, but which are related, at least, to social adaptability. More work is necessary, however, before the function measured by the test can be specifically designated "social intelligence."

The Kohs Ethical Discrimination Test (p. 151) has not been validated by correlation with other measures. Since this test involves reading and language comprehension to a considerable degree, it is probably as much a test of ethical knowledge or ethical vocabulary, or of docility and agreement with conventional moral standards, as it is of ethical behavior. It is doubtless true, however, that considerable insight might be had into a child's attitudes toward conduct and his concepts of right and wrong behavior from his test score.

In the construction of their tests for measuring character traits, Hartshorne and May have given especial attention to the question of validity (26). It has been pointed out by these authors that the first step in securing a valid test is to make a detailed analysis of the behavior which one wishes to measure. The next step, of course, is to devise test material or test situations which will adequately sample the field of one's interest. This type of validation has been previously discussed in connection with questionnaires (p. 144).

The validation of a completed character test is discussed by Hartshorne and May from two points of view, that of (1) empirical validity, and (2) theoretical validity. Empirical validity is the usual method employed in mental tests; it refers to the correlation of a given test with other objective measures (external criteria) of the behavior studied. Ratings by teachers was the first criterion used by Hartshorne and May. For example, 480 children in Grades 5 to 8 were rated for honesty on an eleven-point scale by their teachers, and these ratings were correlated with cheating scores made on the tests. When the correlations between ratings and scores were corrected for attenuation, and all rating scales discarded on which only one or two steps were used, the correlation of ratings for honesty and honesty scores made in classroom tests was approximately .40. This validity correlation is but slightly less than the correlation of .50 between general intelligence test scores and ratings for intelligence in the same group. Neither of these figures, to be sure, is particularly high, but both are fairly adequate considering the imperfect nature of the criteria. To attempt to validate one set of inadequate measures against another set just as inadequate is really going in a circle. About all that one can say with assurance is that both the tests and the ratings are recognizing and measuring to a fair degree the same kind of behavior.

Two other empirical methods were used by Hartshorne and May in evaluating their honesty test, *viz.*, validation by records, and validation by confession. To illustrate the first, in one private school of 146 pupils, records were kept of whether a child was above or below average in honesty, as well as in other desirable forms of conduct. Of the seven pupils noted as below average in honesty, three cheated on the I.E.R. tests; while of the 139 marked above average, only 18 per cent. cheated. This result suggests that the tests were getting at the same kind of dishonest behavior noted in the records. In validation by confession, the fact of whether or not a child cheated on a test was checked against a questionnaire given afterward, in which the child was asked directly whether he had gotten help, and whether he regarded such action as dishonest. In a population of 2,141 children, 44 per cent. were apparently dishonest—used a key unfairly, or got help at home. Of this group, 89 per cent. answered the question about getting aid unfairly, 82 per cent. agreeing that to copy from a key was cheating. Of the 56 per cent. *not* cheating—

or at least not detected—93 per cent. answered the question about getting aid unfairly, all saying that such practice was 'dishonest. Assuming that these answers reflect the pupils' real opinions, Harts-horne and May comment that "In about 90 per cent. of the subjects measured, the scores represent not only the fact of deception, or falsification, but also the attitude of deception, or the feeling that the act is a genuine act of cheating."

Empirical validity is not wholly satisfactory, because the criteria which we set up are never entirely adequate. For this reason, Harts-horne and May calculated what they called the "theoretical validity" of their tests (26 [140-144]). This is defined as the degree to which a battery of tests truly represents the results to be expected from a very large number of measured determinations of the behavior, *viz.*, from an infinite number of tests of the same sort. The problem is largely a statistical and technical one, and the reader is referred to May's discussion of the topic (26 [118-129]). An example of the method will suffice for purposes of illustration. In four tests designed to reveal classroom dishonesty (*e.g.*, I.E.R., Speed, Coördination, and Puzzles) the average intercorrelation was found to be .39. The correlation of these four tests taken together, with four equally good tests, will be .72 by the Spearman-Brown prophecy formula (22), and this *r* may be taken as the reliability coefficient of the battery as a whole. The correlation of this battery with "true" measures of itself (22) (theoretical "true" criterion) is now found by taking the square root of the reliability coefficient, *i.e.*,  $\sqrt{.72}$ . This gives .85 as the theoretical validity of the battery—of how well it represents an infinite number of similar measures of the behavior in question. Other illustrations of this concept of theoretical validity and of its use in the evaluation of a test battery may be found in Hartshorne and May's work (27, 28).

#### BIBLIOGRAPHY

1. ALLPORT, F. H., *Social Psychology*, Houghton Mifflin Company, New York, 1924.
2. ALLPORT, G. W., "A Test for Ascendance-Submission," *Journal Ab-normal and Social Psychology*, 23:118-136, 1928.
3. BARRETT, M., "The Order of Merit Method and the Method of Paired Comparisons," *Journal Philosophy*, 10:382-384, 1913.
4. BERNREUTER, R. G., *The Personality Inventory*, Stanford University Press, California, 1931.

5. BILLS, M. A., "A Method for Classifying the Jobs and Rating the Efficiency of Clerical Workers," *Journal Personnel Research*, 1:384-393, 1923.
6. BOYCE, A. C., "Methods for Measuring Teachers' Efficiency," *14th Year-book, National Society for the Study of Education*, Part II, 1915.
7. BRADSHAW, F. F., "American Council on Education Rating Scale," *Archives Psychology*, 119, 1930.
8. BRIDGES, J. W., "Emotional Instability of College Students," *Journal Abnormal and Social Psychology*, 22:227-234, 1927.
9. CADY, V. M., "The Estimation of Juvenile Incurability," *Journal Delinquency Monographs*, 2, 1923.
10. CHAMBERS, O. R., "Character Trait Tests and the Prognosis of College Achievement," *Journal Abnormal and Social Psychology*, 20:303-311, 1925.
11. CHAMBERS, O. R., "A Method of Measuring the Emotional Maturity of Children," *Pedagogical Seminary and Journal Genetic Psychology*, 32:637-647, 1925.
12. CONKLIN, E. S., "The Determination of Normal Extrovert-Introvert Interest Differences," *Pedagogical Seminary and Journal Genetic Psychology*, 34:28-37, 1927.
13. COWDERY, K. M., "Measurement of Professional Attitudes," *Journal Personnel Research*, 5:131-141, 1926.
14. DOWNEY, J. E., "Some Volitional Patterns Revealed by the Will-Profile," *Journal Experimental Psychology*, 3:281-301, 1920.
15. DOWNEY, J. E., AND UHRBROCK, R. S., "Reliability of the Group Will-Temperament Tests," *Journal Educational Psychology*, 19:26-39, 1927.
16. FLEMMING, E. C., "The Predictive Value of Certain Tests of Emotional Stability as Applied to College Freshmen," *Archives Psychology*, 96, 1928.
17. FRANZ, S. I., *Handbook of Mental Examination Methods*, The Macmillan Company, New York, 1919.
18. FREYD, M., "The Graphic Rating Scale," *Journal Educational Psychology*, 14:83-102, 1923.
19. FREYD, M., "Introverts and Extroverts," *Psychological Review*, 31:74-87, 1924.
20. FREYD, M., "The Personalities of the Socially and Mechanically Minded," *Psychological Monographs*, 33, 1924.
21. GALTON, FRANCIS, *Inquiries into Human Faculty and Its Development*, The Macmillan Company, New York, 1883.
22. GARRETT, H. E., *Statistics in Psychology and Education*, Longmans, Green and Company, New York, 1926.
23. GARRETT, H. E., AND KELLOGG, W. N., "The Relation of Physical Constitution to General Intelligence, Social Intelligence and Emotional Instability," *Journal Experimental Psychology*, 11:113-129, 1928.
24. GARRETSON, O. K., *Relationships between Expressed Preferences and*

- Curricular Abilities of Ninth-Grade Boys*, Teachers College, Columbia University, Contributions to Education, 386, 1930.
25. HART, H. N., "A Test of Social Attitudes and Interests," *University of Iowa Studies in Child Welfare*, II, No. 4, 1923.
  26. HARTSHORNE, H., AND MAY, M., *Studies in Deceit*, The Macmillan Company, New York, 1928.
  27. HARTSHORNE, H., MAY, M., AND MALLER, J., *Studies in Service and Self-Control*, The Macmillan Company, New York, 1929.
  28. HARTSHORNE, H., MAY, M., AND SHUTTLEWORTH, F. K., *Studies in the Organization of Character*, The Macmillan Company, New York, 1930.
  29. HEIDBREDER, E., "Introversion and Extroversion in Men and Women," *Journal Abnormal and Social Psychology*, 22:52-61, 1927.
  30. HOITSMA, R. K., "The Reliability and Relationships of the Colgate Mental Hygiene Test," *Journal Applied Psychology*, 9:293-303, 1925.
  31. HOLLINGWORTH, H. L., *Psychology of Functional Neuroses*, D. Appleton and Company, New York, 1920.
  32. HOUSE, S. D., "A Mental Hygiene Inventory," *Archives Psychology*, 88, 1927.
  33. HUBBARD, R. M., "A Measure of Mechanical Interests," *Journal Genetic Psychology*, 35:229-254, 1928.
  34. HUNT, T., "The Measurement of Social Intelligence," *Journal Applied Psychology*, 12:317-334, 1928.
  35. KELLEY, T. L., *Statistical Method*, The Macmillan Company, New York, 1923.
  36. KORNHAUSER, A. W., "Reliability of Average Ratings," *Journal Personnel Research*, 5:309-317, 1926.
  37. LAIRD, D. A., "Detecting Abnormal Behavior," *Journal Abnormal and Social Psychology*, 20:128-141, 1925.
  38. MARSTON, L. R., "The Emotions of Young Children," *University of Iowa Studies in Child Welfare*, III, 1925.
  39. MATHEWS, E., "A Study of Emotional Stability in Children," *Journal Delinquency*, 8:1-40, 1923.
  40. MCGEOCH, J. A., AND WHITELY, P. C., "The Reliability of the Pressey X-O Test for Investigating the Emotions," *Pedagogical Seminary and Journal Genetic Psychology*, 34:255-270, 1927.
  41. MEIER, N. C., "A Study of the Downey Test by the Method of Estimates," *Journal Educational Psychology*, 14:385-395, 1923.
  42. PATERSON, D. G., "The Scott Company Graphic Rating Scale," *Journal Personnel Research*, 1:361-376, 1923.
  43. PATERSON, D. G., "Methods of Rating Human Qualities," *Annals American Academy of Political and Social Science*, No. 199, 110: 81-93, 1923.
  44. PEARSON, KARL, "On the Relationship of Intelligence to Size and Shape of Head and to Other Physical and Mental Characters," *Biometrika*, 5:105-146, 1906-1907.
  45. *Personnel System of the U. S. Army*, vol. II, Washington, D. C., 1919.

46. PRESSEY, S. L., "A Group Scale for Investigating the Emotions," *Journal Abnormal and Social Psychology*, 16:55-64, 1921.
47. RAUBENHEIMER, A. S., "An Experimental Study of Some Behavior Traits of the Potentially Delinquent Boy," *Psychological Monographs*, 34:6, 1925.
48. REAM, M. J., *Ability to Sell: Its Relation to Certain Aspects of Personality and Experience*, The Williams and Wilkins Company, Baltimore, Maryland, 1924.
49. RUCH, G. M., "A Preliminary Study of the Correlations between Estimates of the Volitional Traits and the Results of the Downey 'Will Profile,'" *Journal Applied Psychology*, 5:159-162, 1921.
50. RUCH, G. M., AND DEL MANZO, M. C., "The Downey Will-Temperament Group Test: Analysis of Its Reliability and Validity," *Journal Applied Psychology*, 7:65-76, 1923.
51. RUGG, H. O., "Is the Rating of Human Character Practicable?" *Journal Educational Psychology*, 12:425-438, 485-501, 1921; 13:30-42, 81-93, 1922.
52. SCOTT, W. D., AND CLOTHIER, R. C., *Personnel Management*, A. W. Shaw and Company, Chicago, 1923.
53. SHEN, E., "The Reliability Coefficient of Personal Ratings," *Journal Educational Psychology*, 16:232-236, 1926.
54. SLAWSON, J., "Psychoneurotic Responses of Delinquent Boys," *Journal Abnormal and Social Psychology*, 20:261-281, 1925.
55. STRONG, E. K., "The Interest Tests for Personnel Managers," *Journal Personnel Research*, 5:194-203, 1926-1927.
56. STRONG, E. K., "Procedure for Scoring an Interest Test," *Psychological Clinic*, 19:63-72, 1930.
57. STRONG, E. K., "Diagnostic Value of the Vocational Interest Test," *Educational Record*, 10:59-68, 1929.
58. STRONG, E. K., *Manual for Vocational Interest Blank*, Stanford University Press, California, 1930.
59. SYMONDS, P. M., "On the Loss of Reliability in Ratings Due to Coarseness of the Scale," *Journal Experimental Psychology*, 7:456-460, 1924.
60. SYMONDS, P. M., *Diagnosing Personality and Conduct*, The Century Company, New York, 1931.
61. SYMONDS, P. M., "A Social Attitudes Questionnaire," *Journal Educational Psychology*, 16:316-322, 1925.
62. TERMAN, L. M., *Genetic Studies of Genius*, I, Stanford University Press, California, 1925.
63. THORNDIKE, E. L., "A Constant Error in Psychological Ratings," *Journal Applied Psychology*, 4:25-29, 1920.
64. THURSTONE, L. L., "An Experimental Study of Nationality Preferences," *Journal General Psychology*, 1:405-425, 1928.
65. THURSTONE, L. L., AND CHAVE, E. J., *The Measurement of Attitude*, University of Chicago Press, Chicago, 1929.

66. THURSTONE, L. L., "Attitudes Can Be Measured," *American Journal Sociology*, 33:529-554, 1928.
67. THURSTONE, L. L., AND THURSTONE, T. G., "A Neurotic Inventory," *Journal Social Psychology*, 1:3-30, 1930.
68. THURSTONE, L. L., AND THURSTONE, T. G., *Manual for Using the Personality Schedule*, University of Chicago Press, Chicago, 1930.
69. UHRBROCK, R. S., *An Analysis of the Downey Will-Temperament Tests*, Teachers College, Columbia University, Contributions to Education, 296, 1928.
70. VOELKER, P. F., *The Function of Ideals and Attitudes in Social Education*, Teachers College, Columbia University, Contributions to Education, 112, 1921.
71. WATSON, G. B., *The Measurement of Fairmindedness*, Teachers College, Columbia University, Contributions to Education, 176, 1925.
72. WYMAN, J. B., "The Measurement of Interest," *Vocational Guidance Magazine*, 8:54-60, 1929.

## CHAPTER IV

### TESTS IN SPECIAL FIELDS

IN RECENT years there has been developed an increasing number of tests designed to measure aptitude in various special fields. Many of these tests have proved to be exceedingly valuable in discovering the extent of an individual's special "gifts," as well as in predicting his probable success in a vocation wherein certain definite preparation is demanded. To treat fully all of the tests constructed for a specific purpose, or planned to meet some definite vocational need, is a task beyond the limits of this chapter. And in any event it would be unnecessary, as several groups of such tests have been fully described elsewhere. For example, trade tests, which are usually work samples or vocational miniatures, have been discussed by Toops (20) and by Chapman (2); and the construction of test batteries planned to measure vocational aptitudes has been carefully outlined by Hull (8). In Chapter II we have described tests of mechanical ability, and in Chapter III, methods of measuring personality and character traits.

For the purposes of this chapter we have divided tests in special fields into two groups. In the first group are those tests designed to estimate ability in the special and well-defined fields of art and music. In the second group are tests concerned specifically with an individual's preparation for, and probable success in, some vocation, such as medicine, law or clerical work. Our first group of tests is probably measuring native endowment to a greater degree than the second, in which training, developed interests and environmental influences admittedly loom large. However, there is certainly much overlapping between the two groups and any clear-cut distinction in terms of heredity or environment is unwarranted.

In the following sections representative groups of tests selected from various special fields will be described, and the validity of such batteries will be considered. Some typical results obtained with such tests will be found in Chapter V.



## REPRESENTATIVE TESTS IN SPECIAL FIELDS

## MUSIC

## 1. Measures of Musical Talent, by Carl E. Seashore

*Date:* 1919.

*Publisher:* The C. H. Stoelting Company, Chicago; also Columbia Phonograph Company, Education Department, New York, New York.

*Designed for:* Grades 5 to 8, and adults.

*Contents:* Six double-disk records which can be played on any standard phonograph. Records are measures of

- (1) Pitch discrimination
- (2) Intensity
- (3) Time
- (4) Consonance
- (5) Tonal memory
- (6) Rhythm.

*Scoring:* Per cent. correct of total responses. Relative attainment may be studied by means of graphs.

*Norms:* Percentile tables are given in Manual for Grade 5, Grade 8, and for adults.

*Reliability:* For different parts of the test, reliabilities vary from .35 (consonance) to .70 (pitch),  $N = 100$ . Reliability coefficients will depend greatly upon the age and training of subjects, as well as upon the training of the examiner.

## 2. Music Test, by H. E. Hutchinson and L. W. Pressey

*Date:* 1924.

*Publisher:* Public School Publishing Company, Bloomington, Illinois.

*Designed for:* Grades 7 to 12.

*Contents:* This is a test of ability to read music silently, and to recognize musical scores from known songs and operas. There are six groups of musical selections. Groups 1 to 5 consist of four and Group 6 of five musical lines from different compositions. After a trial series, the child is required to mark each musical line by a number which designates his choice. Eight to ten titles are given from which selection is to be made.

*Scoring:* Total number of correct responses, maximum score being twenty-five.

*Norms:* Tentative norms (medians) for Grades 7, 8, 9, 10, 11, 12 are given in the Manual.

*Reliability:* Not reported.

## 3. Sight-Singing Test, by E. K. Hillbrand

*Date:* 1923.

*Publisher:* World Book Company, Yonkers, New York.

*Designed for:* Grades 4, 5, 6.

*Contents:* This is an individual test designed to measure the child's abil-

ity in the oral sight-reading of vocal music. There are six simple musical selections (words and music) which the pupil is instructed to sing.

*Scoring:* Various kinds of errors are noted and recorded by the experimenter, *viz.*:

- (1) Notes wrongly pitched
- (2) Transpositions
- (3) Flatting
- (4) Sharping
- (5) Omission of notes
- (6) Errors in time.

In addition the following irregularities, *viz.*:

- (7) Extra notes
- (8) Repetitions
- (9) Hesitations

are noted but not counted as errors.

*Norms:* Tentative grade norms in terms of errors are given in Manual.

*Reliability:* Not reported.

#### **Test of Music Information and Appreciation, by J. Kwalwasser**

*Date:* 1927.

*Publisher:* Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

*Designed for:* High school and college.

*Contents:* Three parts as follows:

##### **I. History and Biography**

- (1) Classification of artists
- (2) Nationality of composers
- (3) Composers of famous compositions
- (4) Classification of composers by types of compositions
- (5) General knowledge of composers and compositions

##### **II. Instrumentation**

- (1) Production of tones on orchestral instruments
- (2) Classification of orchestral instruments
- (3) General knowledge of instrumentation

##### **III. Musical Form**

- (1) General knowledge of music structure and form.

*Scoring:* Total scores translated into percentiles.

*Norms:* Not given.

*Reliability:* .70 to .72 (N not given).

#### **Test of Musical Accomplishment, by J. Kwalwasser and G. M. Ruch**

*Date:* 1924, revised 1927.

*Publisher:* University of Iowa, Iowa City, Iowa.

*Designed for:* Grades 4 to 12.

*Contents:* This test has ten parts:

- (1) Knowledge of musical symbols and terms
- (2) Recognition of syllable names

- (3) Detection of pitch errors in a familiar melody
- (4) Detection of time errors in a familiar melody
- (5) Recognition of pitch names
- (6) Knowledge of time signatures
- (7) Knowledge of key signatures
- (8) Knowledge of note values
- (9) Knowledge of rest values
- (10) Recognition of familiar melodies from notation.

*Scoring:* Number right. Each exercise is weighted on a scale of 1 to 5. Total score is sum of scores on all of the exercises.

*Norms:* Average total scores are given for Grades 4 to 12. Decile scores are given in Manual.

*Reliability:* .97 (split-half),  $N = 167$  pupils in Grades 6, 8, 10, 12. The reliabilities of the sub-tests in the same group range from .70 to .97.

#### ART

##### 1. Art Judgment Test, by Norman C. Meier and Carl E. Seashore

*Date:* 1929.

*Publisher:* Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.

*Designed for:* High schools and art schools.

*Contents:* This test consists of a book containing 125 pages of pictures. These pictures are printed in phototone and resemble etchings. They are arranged in pairs, the two pictures in each pair being alike except in one respect. S is told in what respect the two pictures differ, and is instructed to decide which is the better of the two (more pleasing, more artistic, more satisfying). If the left-hand picture is preferred, L is circled; if the right-hand, R is circled.

*Scoring:* The number right out of 125 comparisons determines the score. Scoring is done by means of a stencil.

*Norms:* Percentile norms have been established for Grades 7, 8, 9, 10, 11 and 12 based upon 1,850 students. Qualitative descriptions are given, also, for scores falling within each quarter of the distribution.

*Reliability:* .71 (repetition) for sixty-nine undergraduates; .71 (split-half) for 100 intermediate art students and high-school students; and .85 (split-half) in another group of seventy-seven art and high-school students.

##### 2. Art Test, by Margaret McAdory

*Date:* 1929.

*Publisher:* Bureau of Publications, Teachers College, Columbia University, New York.

*Designed for:* Elementary and high school pupils and for adults.

*Contents:* These tests consist of seventy-two plates, in black and white, and in color. Each plate presents four illustrations of the same subject (designated A, B, C and D) treated in slightly different ways. The illus-

trations deal with articles of furniture, utensils, textiles and clothing, architecture, painting and other graphic arts, portraits, *etc.*

*Scoring:* S is required to indicate his first, second, third, and fourth choices. Total number of correct judgments (out of a total of 288) gives the score. This score may be expressed as a percentage. An analysis is also made of S's judgment with respect to the various kinds of subjects treated in the illustrations.

*Norms:* Means for sixth grade, college freshmen, college seniors, competent critics,  $N = 100$  in each case. Norms are tentative and are based upon the test in its experimental form.

*Reliability:* .93 (split-half),  $N = 100$  adults; .80,  $N = 100$  sixth-grade children.

*Reference:* McADORY, M., *The Construction and Validation of an Art Test*, Teachers College, Columbia University, Contributions to Education, 383, 1929.

**Test in Fundamental Abilities of Visual Art, by A. S. Lewerenz**

*Date:* 1927.

*Publisher:* Southern California Book Depository, Los Angeles, California.

*Designed for:* Third grade to college level.

*Contents:* This test consists of nine separate tests arranged in three groups as follows:

Part I

- (1) Recognition of proportion
- (2) Originality of line drawing

Part II

- (3) Observation of light and shade
- (4) Knowledge of subject matter
- (5) Visual memory of proportion

Part III

- (6) Analysis of problems in cylindrical perspective
- (7) Analysis of problems in parallel perspective
- (8) Analysis of problems in angular perspective
- (9) Recognition of color.

*Scoring:* Total number correct.

*Norms:* For Grades 3 to 12, based upon 1,100 pupils.

*Reliability:* .87 (retest),  $N = 100$  pupils in Grades 3 to 9.

COMMERCIAL AND VOCATIONAL APTITUDE

**Bookkeeping Test, by Fayette H. Elwell and John Guy Fowlkes**

*Date:* 1928.

*Publisher:* World Book Company, Yonkers, New York.

*Purpose:* "To furnish a reliable, valid, and comparable measure of achievement in bookkeeping for use in high schools and business colleges."

*Contents:* This test consists of nine parts and covers the five main di-

visions of bookkeeping, *viz.*: general theory, journalizing, classification, adjusting entries and closing the ledger, and statements. There are two tests, each consisting of two forms, A and B.

*Scoring*: Score is number right in all parts but five and nine, where number right is divided by two. Total score is the sum of separate parts.

*Norms*: Means and S.D.'s for Tests 1 and 2 are supplied, based in each instance upon approximately 250 cases. The subjects were high-school students. Percentile distributions for the same groups are also given.

*Reliability*: For Test 1, the reliability coefficient is .82 ( $N = 256$ ); for Test 2, .87 ( $N = 226$ ). The subjects were high-school students.

## 2. Clothing Test, by Florence D. Frear and Warren W. Coxé

*Date*: 1929.

*Publisher*: Public School Publishing Company, Bloomington, Illinois.

*Purpose*: To measure the knowledge of clothing possessed by high-school girls who are taking or have completed courses in elementary dressmaking.

*Contents*: There are five parts to the test as follows:

Part I. Fundamentals of construction

Part II. Care and repair of clothing

Part III. Hygiene of clothing

Part IV. Appropriateness of clothing

Part V. Economics of clothing.

*Scoring*: Each sub-test is scored by means of a scoring key, the scores from the individual tests being summed to give the final score.

*Norms*: Norms in terms of total score (medians) are given for one, two and three semesters of work in dressmaking.  $N$  is not stated.

*Reliability*: Not reported.

## 3. Examination in Clerical Work, by L. L. Thurstone

*Date*: 1922.

*Publisher*: World Book Company, Yonkers, New York.

*Purpose*: To measure aptitude for clerical and general office work.

*Contents*: This is a paper-and-pencil group test which consists of eight parts, or sub-tests, as follows: (1) checking for arithmetic errors; (2) locating misspelled words; (3) cancellation; (4) code learning; (5) alphabetizing, *i.e.*, writing a list of thirty-seven names in ten groups in *a b c* order; (6) directions test involving classification of life insurance policies; (7) arithmetic problems; (8) matching eleven proverbs against eleven other proverbs so that two proverbs in each pair have the same meaning.

*Scoring*: Scoring is in terms of errors and time, from which may be calculated an accuracy rating, a speed rating and a combined speed and accuracy rating.

*Norms*: Not given.

*Reliability*: Not given.

**1. Home Economics Test, by Edna M. Engle and John L. Stenquist**

*Date:* 1931.

*Publisher:* World Book Company, Yonkers, New York.

*Purpose:* "To provide an objective measure of a pupil's knowledge in the several fields of Home Economics."

*Designed for:* Grades 5 to 10.

*Contents:* There are three separate tests covering (1) Foods and Cookery, (2) Clothing and Textiles and (3) Household Management. The first test deals with selection of foods, nutrition, marketing, table service and etiquette, planning, preparation and serving of meals. The second test covers planning and construction of garments, study of textiles, selection, care and repair of clothing and factors to be considered in buying clothing. The third test deals with the study, care and furnishing of the home, budgeting, and family relationships. Each test consists of four parts, and has two forms, A and B.

*Scoring:* Scores on separate tests are number R, or R - W, the correct answers being given in a key.

*Norms:* Age and grade norms for each of the three tests are given. According to the authors, norms are based upon 15,000 cases for Foods and Cookery, and for Clothing and Textiles, and 10,000 for Household Management. Subjects were public school children.

*Reliability:* Foods and Cookery, .92 to .94 in Grades 5 to 8 (N = 110 to 117); Clothing and Textiles, .93 to .96 in Grades 5 to 8 (N = 110 to 149); Household Management, .85 to .96 in Grades 6 to 8 (N = 103 to 125).

**Stenogauge Test, by E. J. Benge**

*Date:* 1923.

*Publisher:* Stenogauge Company, 3136 North 24th Street, Philadelphia, Pa.

*Purpose:* To measure aptitude for stenography.

*Contents:* This test consists of four parts. (1) Dictation: a letter, several hundred words in length, is dictated to the applicant at the maximum rate at which he can record it. Time in seconds is noted. (2) Transcription rate: time required by candidate to transcribe his notes is recorded in seconds. (3) Transcription accuracy: a percentage found by dividing number of words transcribed correctly, by number of words in the letter. (4) Spelling: applicant is given a list of fifty words, half of which are incorrectly spelled. He is to check those which are misspelled.

*Scoring:* Total score is found by adding together the scores on the four tests: number of words dictated per minute, number of transcriptions per minute, the percentage of total transcriptions done correctly, and the percentage of the list of fifty words checked correctly.

*Norms:* Percentiles based upon approximately 500 cases are given in the Manual.

*Reliability:* Not reported.

*Reference:* FREYD, MAX, "Selection of Typists and Stenographers: Information on Available Tests," *Journal Personnel Research*, 5:490-510, 1926-1927.

6. Test for Ability to Sell, by F. A. Moss, Herbert Wyle, Wm. Loman and Wm. Middleton

*Date:* 1929.

*Publisher:* Center for Psychological Service, Washington, D. C.

*Purpose:* To estimate ability to sell merchandise.

*Contents:* This test is a paper-and-pencil group test. It contains six parts or sub-tests as follows:

- Test 1. Judgment in selling situations
- Test 2. Memory for names and faces
- Test 3. Observation of behavior
- Test 4. Learning selling points in merchandise
- Test 5. Following store directions
- Test 6. Selling problems.

*Scoring:* Scoring is in terms of points, the final score being the sum of the points earned on the separate tests.

*Norms:* Median scores are supplied for groups of different educational levels, e.g., seventh grade through college. Average score of the twenty-five best salespersons in one large department store chain was 112, the average of the poorest salespersons, 70.5.

*Reliability:* .91 (N not given).

7. Typewriting Test (Stenographic Proficiency Tests), by E. G. Blackstone

*Date:* 1923.

*Publisher:* World Book Company, Yonkers, New York.

*Purpose:* To measure ability in typewriting.

*Contents:* A business letter of approximately 200 words to be typed by the students. There are five forms, in order that the test may be given at intervals without repetition.

*Scoring:* Speed and accuracy are combined into the following final score:

$$\text{Score} = \frac{\text{strokes per minute} \times 10}{\text{errors} + 10}$$

*Norms:* Based upon 2,188 cases.

*Reliability:* .93, on average, for groups of pupils with twenty months' instruction. P.E. score = 5.8, N = 105.

### PROFESSIONAL APTITUDE

1. Law Aptitude Examination, by M. L. Ferson and G. D. Stoddard

*Date:* 1925, 1927.

*Publisher:* West Publishing Company, St. Paul, Minnesota.

*Purpose:* Designed for law students or law school candidates.

*Contents:* This test consists of four parts or sub-tests. These demand

accurate recall; comprehension of difficult reading matter; reasoning by analogy; reasoning by analysis; and skill in pure logic of the syllogistic sort.

*Scoring* The scores in Parts 2 and 3 are number right; in Part 1-B each correct answer is weighted 2; and in Part 4 each correct answer is weighted 3.

*Norms*: Medians, twenty-fifth and seventy-fifth percentiles based upon results from eight law schools.

*Reliability*: Not reported.

*Reference*: FERSON, M. L., AND STODDARD, G. D., "Law Aptitude," *American Law School Review*, 6:78-81, 1927.

## 2. Prognosis Test of Teaching Ability, by Warren W. Coxe and Jacob S. Orleans

*Date*: 1930.

*Publisher*: World Book Company, Yonkers, New York.

*Purpose*: To aid teacher-training institutions in selecting their applicants.

*Contents*: This test consists of five parts. Part I: General Information, including literary, scientific and historical facts. Part II: Professional Interest; a series of questions on teachers' duties and classroom practice. Part III: Lessons in Education; comprises eight brief lessons in educational psychology, educational measurement, principles of teaching, evaluation of objectives, etc. Part IV: Reading Comprehension; consists of seven reading tests based upon material taken from textbooks in education. Part V: Problems in Education; presents six educational problems, with questions upon each, designed to test the students' knowledge and judgment. The authors suggest that in setting up standards for entrance to normal schools as well as in classifying those students who enter, the high school record, work habits, and general intelligence of the student should be considered as well as the more professional aspects of his training which are measured by the present test.

*Scoring*: A weighted scoring method, described in the Manual, is employed.

*Norms*: A table is given showing the probability of success in teacher training (first year) from standing in the Prognosis Test.

*Reliability*: Not reported.

## 3. Scholastic Aptitude Tests for Medical Schools, by F. A. Moss, O. B. Hunter and H. F. Hubbard

*Date*: 1929.

*Publisher*: Center for Psychological Service, Washington, D. C.

*Purpose*: To predict probable success in medicine of pre-medical students and first-year medical students.

*Contents*: This is a paper-and-pencil test. It consists of a booklet containing six tests as follows:



- Test 1. Scientific vocabulary
- Test 2. Pre-medical information
- Test 3. Visual memory
- Test 4. Memory for content
- Test 5. Comprehension and retention
- Test 6. Understanding of printed material.

*Scoring:* Number of points earned out of a possible total of 250.

*Norms:* Percentage distributions of school grades made by freshmen medical students are given for each of the ten deciles of test scores; also median freshman grades and quartiles for each score decile. To illustrate, those students whose aptitude test scores placed them in the highest 10 per cent. had school records as follows: 10 per cent, ninety or above; 83 per cent., eighty to eighty-nine, and 7 per cent., seventy-five to seventy-nine. Results are based upon more than thirty medical schools.

*Reliability:* .93 (N not given).

*Reference:* Moss, F. A., "Scholastic Aptitude Tests for Medical Students," *Journal Association American Medical Colleges*, 5:90-110, 1930.

#### 4. Stanford Educational Aptitude Tests, by M. B. Jensen

*Date:* 1928.

*Publisher:* Stanford University Press, Stanford University, California.

*Purpose:* To offer a comparative estimate of a student's fitness for teaching, educational research or school administration.

*Contents:* This test is really a questionnaire, in that it calls for decisions and opinions based upon judgment and experience in educational work. There are three parts:

- Position preference ratings
- Discipline case problems
- High school activities.

In each part of the test various problems are presented to which alternative solutions are suggested. Subjects are asked to select one alternative and to denote their confidence in this judgment on a four-point scale.

*Scoring:* A weighted scoring key is used, the final score on the test as a whole serving to designate the field in which an individual will probably do his best work.

*Norms:* Tables are provided from which the probability of an individual's success in each of the three fields covered by the test may be obtained.

*Reliability:* Differential scores in the battery, *i.e.*, the amount by which the score in one field exceeds the score in another, have reliability coefficients from .85 to .94, N = 50 in each case.

*Reference:* JENSEN, M. B., "Objective Differentiation between Three

Groups in Education, Teachers, Research Workers, and Administrators," *Genetic Psychology Monographs*, 3:333-454, 1928.

. Stanford Scientific Aptitude Test, by D. L. Zyve

*Date:* 1929.

*Publisher:* Stanford University Press, Stanford University, California.

*Designed for:* College students.

*Contents:* Eleven exercises, questions, problems upon the following topics:

- (1) Experimental bent
- (2) Clarity of definition
- (3) Suspended *vs.* snap judgment
- (4) Reasoning
- (5) Inconsistencies
- (6) Fallacies
- (7) Induction, deduction and generalization
- (8) Caution and thoroughness
- (9) Discrimination of values in selecting and arranging experimental data
- (10) Accuracy of interpretation
- (11) Accuracy of Observation.

*Scores:* Each exercise is weighted (weights ranging from 2 to 7). The method of scoring the separate exercises is given in the Manual. The final score is total of scores on sub-tests.

*Norms.* Given for an unselected group of 324 students, and for forty-seven seniors and graduates of non-scientific groups.

*Reliability:* .93, estimated for 377 students, freshmen to graduates.

Vocational Guidance Tests for Engineers, by L. L. Thurstone

*Date:* 1922.

*Publisher:* World Book Company, Yonkers, New York.

*Purpose:* To determine the probable success in an engineering course of high-school seniors and college freshmen.

*Contents:* There are five tests in this battery, one in each of the following subjects: Arithmetic, Algebra, Geometry, Physics, Technical Information. The first four tests consist of problems and exercises, some of which require compass and straight edge. The fifth test contains 100 questions covering general technical information such as a boy in high school might acquire through his own initiative and interest.

*Scoring:* Scores on each test are in terms of points earned, answers being scored R or W.

*Norms:* Point scores may be translated into percentage ratings from which a candidate can estimate his relative position among engineering freshmen. Percentile tables are based upon results from forty-three engineering schools and colleges. A subject may also compare his ratings on the five tests with the averages made by various engineering colleges.

*Reliability:* Not reported.

## VALIDITY

## 1. Music and Art Tests

The validity of many tests in the field of music or art depends directly upon their methods of construction and upon their standardization. Aptitude in art or music cannot be measured in the same way as can performance in arithmetic or in spelling. There are no objective standards which are "right" or "wrong," and the most acceptable criterion against which to evaluate æsthetic judgment or artistic accomplishment is the opinion of acknowledged experts in the field. Many of the tests described on p. 168 have employed this "expert opinion" method of validation. Hillbrand's Sight Singing Test, Hutchinson and Pressey's Music Test, and Kwalwasser's Test of Music Information and Appreciation, all depend for their validity upon the acceptability to musical experts of the material which they include. The authors of the Kwalwasser-Ruch Test of Musical Accomplishment drew their material from the recommendations made by the Music Supervisors' National Conference (16). The "correct" answers in the McAdory Art Test were determined by a consensus of 100 competent judges (12), while the material included in the Lewerenz Test of the Fundamental Abilities in Visual Art was validated on the basis of ratings made by qualified experts in art (11). The final 125 drawings in the Meier-Seashore Art Judgment Test (13) were selected from a group of approximately 600 drawings. All of these final test items were appraised by twenty-five experts for their suitability as test material, as well as for their acceptability as examples of well-recognized æsthetic principles.

The Seashore Measures of Musical Talent (17) are good illustrations of special ability tests which can be evaluated against objective standards. The Seashore records are *bona fide* measures of pitch, intensity, time, *etc.*, since in the Sense of Pitch Test, for instance, the second of the two tones presented is physically higher or lower than the first. In the Sense of Intensity Test, the second of the two intervals is temporally longer or shorter than the first. Only in the Sense of Consonance Test does validity depend upon the judgment of musical experts. Several studies have been made of the relationship of the Seashore tests to musical ability as judged by teachers of music. Brown (1), for example, studied the agreement between the averages of two ratings for "musical capacity" made by the

teacher of music upon 105 high-school students, and the records of these students made on the Seashore tests. The second rating was taken after an interval of four months. Correlations were low, ranging from .11 (Seashore Intensity and ratings) to .41 (Seashore Tonal Memory and ratings). The correlation between the ratings and the whole Seashore battery of tests was .38. Seashore's tests are probably much better measures of musical aptitude than the correlations quoted might suggest. The tests do not claim to be anything more than measures of one's "ear" for music, and besides, criteria of musical accomplishment are difficult to obtain. Judgments made of students for musical accomplishment, for instance, will almost inevitably reflect many factors other than musical aptitude.

The validity of tests of artistic ability may be estimated by their agreement with outside criteria. Lewerenz (10), for example, gave his test of Fundamental Abilities of Visual Art to forty-two students who five months later received grades for work accomplished in art. The correlation between the test scores and the grades was .63—a surprisingly high agreement, considering the size of the group and the probable fallibility of the criterion. Meier and Seashore (13) have employed several ingenious checks upon the authenticity of their Art Judgment Test. Test scores made by individuals trained in art were compared with the scores made by those of little or no training. This comparison showed progressive increases which paralleled closely the amount of training received. These authors stress the fact that some of the untrained group made scores as high as those of the most highly trained. This they take to be evidence that the test does not depend primarily upon training, with the further implication that artistic ability is a special or native "gift." The relation of the Art Judgment Test to measures of general intelligence is low, correlations ranging from .28 to —.14 (*N* varied from 39 to 94). These results plus the fact that university professors, untrained in art, score appreciably lower than highly trained students of art, suggest that artistic ability is not merely a matter of abstract intelligence.

## 2. Vocational and Professional Aptitude

Illustrative of those vocational tests, the validity of which depends upon their construction and standardization, are the Blackstone Typewriting Test, the Elwell-Fowlkes Bookkeeping Test, the Stanford

Educational Aptitude Test of Jensen, and the Engle-Stenquist Home Economics Test. The Blackstone Typewriting Test is in reality a work sample, which, like a trade test, requires that the subject perform the very activities in which his aptitude is being estimated. The Elwell-Fowlkes Bookkeeping Test is a consensus of the material offered in first-year bookkeeping courses (4), while the Engle-Stenquist Home Economics Test is based upon items selected from twelve standard textbooks on the subject, and twenty-five courses in home economics (5).

The validity of the Stanford Educational Aptitude Tests (9) depends upon the weighting of the items, which, as in the Strong Vocational Interest Blank (p. 143) have different values for different professional interests. In drawing up his three scales of educational aptitude, Jensen set up three criterion groups, teachers, research workers and administrators. The members of these groups were selected by seven well-known educators. The responses of these criterion groups were then employed as standards in determining the weights to be attached to the items on each scale. If an item was checked by a relatively greater percentage of the teacher group than of the research worker or the administrator groups, such an item was given greater weight on the teacher scale. In the final scales, one for each of the three groups, items are differentially weighted so as to make maximum contributions to each scale.

External criteria are more easily procured for vocational tests than for tests of æsthetic sensibility. This is largely because most of the examinations designed to assay commercial or professional aptitude can be checked against actual performance. Thurstone (18), for example, has calculated the correlation between his clerical tests and the ratings for efficiency assigned to 100 clerical workers in a large insurance office. The correlation between ratings and accuracy of clerical work was .50; between ratings and speed on the test .42; and between ratings and speed and accuracy together .61. When accuracy and speed on the test were combined with schooling and chronological age of the workers, the composite had a correlation of .67 with the ratings.

The Benge Stenogauge Test (7) was employed as one criterion in the selection of twenty-four individuals for the position of stenographer, out of a total of 145 applicants. Nine of the twenty-four chosen were later rated as being excellent workers, twelve as satis-

factory, and only three as failures. The correlation between ratings for efficiency and the Bengé Stenogauge was .44 ( $N=24$ ); and the correlation between ratings assigned by six executives in another company and the Bengé Stenogauge was .89 ( $N=15$ ). In terms of the criteria employed, it appears that these clerical tests are able to predict success fairly accurately considering the small size of the groups.

Moss and his collaborators report a wide gap between the scores made by the best and the poorest salespersons upon their test for Ability to Sell (14). Salespersons were evaluated as best and poorest on the basis of several criteria: amount of sales, number of errors made, merchandise returned, and combined estimates of buyer, floor manager and personnel officer. Moss reports a correlation of .54 between the scores on the tests and these ratings for sales efficiency.

Since most tests of professional aptitude have been designed for use with students planning to enter a given profession, such examinations are really more a method of selecting promising student material than of predicting later success in the actual practice of a profession. Moss in 1930 reported results from his Scholastic Aptitude Test for Medical Schools (15), based upon returns from twenty-two colleges throughout the United States. Freshmen who took the tests were divided into four equal groups (25 per cent. in each group) on the basis of their test scores; and each group was later studied for its success in the first year of the medical course. In the highest 25 per cent., according to test score, only 1 per cent. failed in the freshman year, 8 per cent. receiving grades over 90 per cent. In the next highest quarter, 8 per cent. failed; in the third quarter 16 per cent., and in the lowest quarter 31 per cent. The correlation between test scores and first-year grades in medicine was .59, indicating that the test has considerable predictive value.

The Law Aptitude Examination by Ferson and Stoddard and the Engineering Aptitude Examination of Thurstone have both been validated against first-year performance in professional schools. Ferson and Stoddard (6) report a correlation of .55 ( $N=395$ ) between test scores and first-semester grades in three law courses; and a correlation of .54 between the test and first-year law school grades in the same group. Thurstone (19) found that the correlation between his five aptitude tests and first-year scholarship in engineering ranged from .23 to .42, with an average of .33. These correla-

tions are slightly higher than those between high-school grades in English, geometry, physics, algebra, and chemistry and first-year engineering grades, which average .30. The samples in these studies varied slightly around 6,500. In drawing up norms for his tests, Thurstone found that the probability was 93 in 100 that a student in the upper 25 per cent. of the test score distribution will remain in good standing in the engineering school. The probability that a student in the second quarter of the test score distribution will remain in good standing is 89 in 100. For a student in the third quarter to remain in good standing, the probability is 81 in 100; and for a student in the fourth quarter, 53 in 100.

The Coxe-Orleans Prognostic Test for Teachers has been validated by a careful choice of material based upon an analysis of the factors which should be considered in determining entrance standards for normal schools; and by its agreement with an objective achievement test (3). The achievement test was administered in ten normal schools at the end of the second semester. Its correlation with the prognostic tests, given upon entrance, ranged from .53 to .84, averaging .65. These correlations were higher than the correlations between the achievement criterion and high-school marks which averaged .50; and about equal to the correlation between the criterion and the Terman Group Intelligence Test. Another test which has been carefully validated is the Stanford Scientific Aptitude Test by D. L. Zyve (21). The fidelity of this battery as a measure of scientific aptitude depends, in the first place, upon the validity of its material, which was selected after a careful analysis of what scientific aptitude may be legitimately taken to imply. Secondly, the validity of the test depends upon the experimental weighting of its component parts, which has been done in such a way as to favor those individuals who are known to possess scientific aptitude. The weights of the different parts of the test were found by comparing the responses of a group of fifty research students in physics, chemical and electrical engineering (the criterion group), with the responses of 121 students specializing in law, literature and languages. Each exercise or part of the scientific aptitude test was weighted in accordance with the relative superiority of the scientific group. The correlation within the criterion group between scores on the test and ratings made by professors for scientific aptitude shown in laboratory and classroom was .74, which was raised to .82 when corrected for attenuation.

Scores on the Scientific Aptitude Test made by twenty-one science faculty members were 35 points higher on the average than the scores made by fourteen non-science faculty members. In a group of forty-seven seniors and graduate students engaged in non-scientific studies the correlation between the aptitude test and scholarship was .02, as against a correlation of .51 between the aptitude test and scholastic grades in the criterion (scientific) group. These correlations indicate clearly the greater importance of scientific aptitude (as defined by the tests) upon achievement in courses in science, than upon achievement in courses in other academic subjects.

## BIBLIOGRAPHY

1. BROWN, A. W., "The Reliability and Validity of the Seashore Tests of Musical Talent," *Journal Applied Psychology*, 12:468-476, 1928.
2. CHAPMAN, J. C., *Trade Tests*, Henry Holt and Company, Inc., New York, 1921.
3. *Coxe-Orleans Prognosis Test of Teaching Ability, Manual of Directions*, World Book Company, Yonkers, New York, 1930.
4. *Elwell-Fowlkes Bookkeeping Test, Manual of Directions*, World Book Company, Yonkers, New York, 1929.
5. *Engle-Stenquist Home Economics Test, Manual of Directions*, World Book Company, Yonkers, New York, 1931.
6. FERSON, M. L., AND STODDARD, G. D., "Law Aptitude," *American Law School Review*, 6:78-81, 1927.
7. FREYD, MAX, "Selection of Typists and Stenographers: Information on Available Tests," *Journal Personnel Research*, 5:490-510, 1926-1927.
8. HULL, C. L., *Aptitude Testing*, World Book Company, Yonkers, New York, 1928.
9. JENSEN, M. B., "Objective Differentiation between Three Groups in Education, Teachers, Research Workers, and Administrators," *Genetic Psychology Monographs*, 3:334-454, 1928.
10. LEWERENZ, A. S., "Predicting Ability in Art," *Journal Educational Psychology*, 20:702-704, 1929.
11. LEWERENZ, A. S., *Test in Fundamental Abilities of Visual Art, Manual of Directions*, Southern California Book Depository, Los Angeles, California, 1927.
12. *McAdory Art Test, Manual of Directions*, Bureau of Publications, Teachers College, Columbia University, 1929.
13. *Meier-Seashore Art Judgment Test, Manual of Directions*, Bureau of Educational Research and Service, University of Iowa, Iowa City, 1930.
14. MOSS, F. A., et al., *Test for Ability to Sell, Manual of Directions*, Center for Psychological Service, Washington, D. C., 1929.



15. MOSS, F. A., "Scholastic Aptitude Tests for Medical Students," *Journal Association American Medical Colleges*, 5:90-110, 1930.
16. RUCH, G. M., AND STODDARD, G. D., *Tests and Measurements in High School Instruction*, World Book Company, Yonkers, New York, 1927.
17. SEASHORE, C. E., *The Psychology of Musical Talent*, Silver, Burdett and Company, New York, 1919.
18. THURSTONE, L. L., "A Standardized Test for Office Clerks," *Journal Applied Psychology*, 3:248-251, 1919.
19. THURSTONE, L. L., *Vocational Guidance Test for Engineers, Manual of Directions*, World Book Company, Yonkers, New York, 1922.
20. TOOPS, H. A., *Trade Tests in Education*, Teachers College, Columbia University, Contributions to Education, 115, 1921.
21. ZYVE, D. L., *Stanford Scientific Aptitude Test, Manual of Directions*, Stanford University Press, California, 1930.

## CHAPTER V

### SOME APPLICATIONS OF PSYCHOLOGICAL TESTS

THE aim of the present chapter is to report upon and evaluate some of the typical investigations carried out through the medium of psychological tests. Tests and testing techniques have been employed extensively upon a variety of problems. Comparative studies of the test scores of twins, siblings and foster children have thrown considerable light upon the relative degree to which individual differences may be thought of as inherited, and the extent to which they may be dependent upon training. The application of mental tests to the extremes of the distribution of abilities, *i.e.*, the "genius" and the "feeble-minded," has led to decided revisions in the traditional concepts of gifted and defective individuals, and of their relationship to the general population. A better understanding of the factors contributing to crime and to juvenile delinquency has also been reached through the use of psychological measurement.

Upon the much-discussed questions of sex differences and racial differences, a mass of data has been accumulated with almost every kind of psychological test. In the fields of industry and in business, tests and measurements have gained widespread use both in the selection of workers, and in the fitting of daily tasks to individual differences in temperament and emotional make-up. Education has offered probably the most fertile field for the application of mental tests. The value of general intelligence and achievement tests in schools has already been discussed at some length in Chapter I, and throughout this book references have been made to the place of tests in education. In the following sections, therefore, only incidental references will be made to the applications of tests in this field.

#### HEREDITY AND ENVIRONMENT

##### 1. Family Resemblances

For showing the relative effects of heredity and environment, a comparison of sibling<sup>1</sup> and twin resemblances upon the same tests is

<sup>1</sup> "Sibling" is a general term for either "brother" or "sister."

especially fruitful. Thorndike's study (51) in 1905 was the first *quantitative* investigation of mental resemblances in twins. Thorndike measured fifty pairs of twins upon six mental tests, *viz.*, cancellation (A, a-t and e-r), misspelled words, addition, multiplication and opposites, comparing the correlation between their scores with the correlation of siblings on the same tests. The  $r$ 's for the twins ranged from .70 to .85 with a mean at .78, while the  $r$ 's for the siblings averaged around .30. When the twins were divided into younger twins (nine to eleven) and older twins (twelve to fourteen) the average correlation for the younger twins was .83, and for the older .70. This result indicates that older twins are less alike than younger twins in the functions measured by the tests. Thorndike argued that if resemblances between twin-pairs in mental test abilities are chiefly the result of common training and common environment, the older twins should have been more alike than the younger, especially in those traits much influenced by training. Since his results showed the opposite to be true, he concluded that the chances decidedly favor common heredity as the cause of twin resemblance.

Many careful studies of twins have been made since Thorndike's investigation, with more cases, better technique and more reliable tests; but the results are not very different from his findings. Merri-man (36), who administered the Stanford-Binet to 105 pairs of twins, obtained a correlation of .88 between the scores of like-sex<sup>1</sup> twins from five to nine years of age; and a correlation of .87 for like-sex twins from ten to fifteen years of age. This consistency in twin resemblance suggests a strong basis in heredity since, if anything, the correlation might be expected to rise with the increase in common environmental influences. Lauterbach (33), who gave eight mental tests to 210 twins, reported an average correlation between like-sex twins (probably identical) of .67, and between unlike-sex or non-identical twins of .41. Lauterbach's summary of the results of his mental and physical tests are given in Table VIII. Note that on the whole resemblances are higher in physical than in mental traits. The same differences between non-fraternal and fraternal twins and

<sup>1</sup> There are two kinds of twins, identical and non-identical, or non-fraternal and fraternal. Identical twins are always of the same sex and are probably produced from the fertilization of a single egg. Non-identical, or fraternal twins, may be of same or opposite sex, and are really a case of dual births, being a result of the fertilization of two eggs. Like-sex twins may or may not be identical; unlike-sex twins cannot be identical.

TABLE VIII  
CORRELATION COEFFICIENTS IN VARIOUS TESTS GIVEN TO TWINS  
OF THE SAME AND OPPOSITE SEXES  
(from Lauterbach [33])

Test	Same Sex	Different Sex
I.Q. ....	.77	.56
Reading Quotient . . . . .	.59	.56
Arithmetic, accuracy. . . . .	.69	.35
Arithmetic, speed . . . . .	.70	.39
Memory for digits . . . . .	.40	.25
Handwriting, quality . . . . .	.69	.37
Handwriting, speed. . . . .	.83	.41
Average . . . . .	.67	.41
Cephalic index. . . . .	.67	.59
Weight. . . . .	.89	.50
Height, standing . . . . .	.80	.53
Height, sitting . . . . .	.73	.59
Average... . . . .	.77	.55

between twins and siblings appear in the study of Tallman (48), who tested 199 siblings and 159 pairs of twins, ranging in age from three to twenty, upon the Stanford-Binet. The average difference in I.Q. was 13.14 points for the siblings; 7.37 points for thirty-nine pairs of non-identical (fraternal) twins; and 5.08 for sixty-three pairs of identical (non-fraternal) twins. These studies of twins and siblings indicate definitely the importance of heredity in test performances. They agree (1) that twins are in general more alike than siblings; and (2) that identical twins are more alike than non-identical in tests of mental abilities. It would seem reasonable to attribute much of the difference between the correlations of twins and siblings to the more nearly identical inheritance of the twins.

It is extremely difficult to identify in the resemblances shown by siblings the relative contributions of immediate ancestry (heredity) and of common training. Thorndike (52) has obtained a correlation (corrected for attenuation) of .60 between the scores made by a group of about 1,200 siblings, upon the Institute for Educational Research Tests of Selective and Relational Thinking, Generalization and Organization.<sup>1</sup> All of these children were high-school students. Thorndike suggests that the difference of .08 between his  $r$  of .60 and the  $r$  of .52 established by Pearson (39) for physical resemblances in siblings can be attributed to the equalizing influence of

<sup>1</sup> A battery of fifteen tests of the type usually found in group tests of general intelligence, e.g., arithmetic, opposites, number series completion, analogies, sentence completion, and the like.

the common school environment upon mental test scores. This surmise, which assigns an extremely meager rôle to environment, can be accepted at face value only if one assumes physical and mental abilities in siblings to be inherited in exactly the same way and to the same degree. If an  $r$  of .52 can be taken to represent the *maximum* effect of heredity, then all correlation over and above this amount would necessarily be due to common environment. A study by Willoughby (55) makes it doubtful that environment *always* has so slight an effect, or that its influence is always in the direction of a leveling-off of native differences. Willoughby reported an average correlation of .42 between the scores of 280 siblings, age seven years and up, upon eleven verbal and non-verbal tests selected from Army Alpha and Beta, N.I.T., and the Stanford Achievement Test. The intercorrelations of these tests in his sibling group ranged from .32, for the Army Beta Test of Geometric Forms, to .50 for the Army Alpha Number Series Completion Test. On the whole, the verbal tests gave higher correlations than the non-verbal, but none of the correlations was as high as the  $r = .60$  obtained by Thorndike.

Part of the discrepancy between these two studies can be fairly attributed to differences in the susceptibility of the respective tests to training, as well as to differences in the size and character of the two groups (age, social and cultural background, *etc.*). Variations in method, and in subjects and material, from study to study, make generalization difficult. However, Willoughby's data suggest strongly either (1) that mental abilities are in many cases less closely bound up in hereditary structure than are physical traits; or else (2) that environment tends to accentuate as often as to equalize individual differences in many tested capacities. Both hypotheses are reasonable. In Thorndike's study, the leveling of abilities (*i.e.*, greater sibling resemblance) through common training seems probable. The subjects were all high-school students, and the tests were closely related to the school subjects to which they had all had equivalent exposure. Willoughby's tests in which the siblings showed the most resemblance were verbal tests. The least resemblance was shown in the non-verbal tests, wherein special training and developed interests would make for differences rather than for likenesses.

The degree of resemblance in mental traits between parents and offspring has been recently studied by H. E. Jones (30), who ex-

amined 210 parents and 317 children in a rural New England community. Children between the ages of three and fourteen were given the Stanford-Binet, while the older children as well as the parents were given the Army Alpha. The multiple correlation of children's scores with *both* parents' scores was .59. Correlations with the parent of the *same* sex as the child were not significantly higher than those with the parent of the other sex, the first relationship averaging .56, the second .54. However, the correlation of the intelligence test scores made by children of either sex with the mother's intelligence test score averaged about five points higher than the correlation with the father's score. Jones attributes this difference (which, though unreliable, was consistent) to the fact that the mother is more intimately connected than the father with the child's early training.

Willoughby in the study already described also tested the parents of his group of siblings with the same eleven tests which had been given to the children. His average parent-child correlation was .35. The correlations of the separate tests varied from .17 for the N.I.T. Test of Checking Similarities to .48 for the History and Literature Information Test of the Stanford Achievement Test. These correlations are lower than the parent-child *r*'s reported by Jones, owing probably to the fact that some of Willoughby's tests measured fairly narrow activities, which are more susceptible to special training and special influences than is the Stanford-Binet. It is interesting to note that Jones' parent-child correlations are in close agreement with the early work of Pearson (39). Pearson established the average parent-child correlation and the average sibling correlation in physical traits, such as hair color, eye color, cephalic index, *etc.*, to be close to .50, and this figure is generally accepted as giving a measure of the hereditary influence in physical traits.

## 2. Foster-children

One of the most extensive studies of the mental resemblances of own and adopted children, and of own and foster parents, is that of Burks (9). Burks tested a total of 214 foster children, all of whom were adopted *before* the age of twelve months, and 105 "control" children living with their own parents. These children were all given the Stanford-Binet and the Woodworth-Cady Personal Data Questionnaire (p. 126). In addition, Burks secured a cultural rating of the foster homes, as well as intelligence test scores, occupational status

and other data upon the foster parents. Information regarding the true parents of adopted children was secured from placement records.

The correlation between the mid-parent<sup>1</sup> intelligence score and foster-child M.A. was .20, as against a correlation of .52 between mid-parent intelligence score and own-child M.A. This lesser resemblance of parents and foster-children strongly indicates the influence of hereditary factors, since the youth of the foster-children, when adopted, allowed environmental factors to operate upon them virtually as long as upon own-children. From all of her data, Burks concluded that the total contribution of heredity alone to individual differences in I.Q. is "probably not far from 75 to 80 per cent." She estimated further that the "very best" environment might raise the I.Q. as much as twenty points; while the "very poorest" environment might lower it as much as twenty points. Changes as great as this, however, rarely occur. A general conclusion suggested by this study is that the Stanford-Binet measures mental performances which are determined principally by hereditary constitution.

Freeman, *et al.* (21), have conducted a somewhat similar study as that of Burks upon 401 foster-children, all of whom were adopted before the age of four years. Of the foster-children, 159 were siblings. In addition, thirty-six own-children of the foster-parents were included for comparative study. All of the children were given Stanford-Binet and the Princeton International Group Mental Test; the foster-parents were given the Otis Self-Administering Test and a fairly difficult vocabulary test. Information was collected upon the economic and cultural status of the foster homes, and occupational, educational and other data upon the foster-parents. Probably the most significant finding of this study was that the I.Q.'s of foster-children show a rise of about seven and a half points on the average when these children are retested after several years of residence in the foster home. Children adopted into the better homes made greater gains in I.Q. than those adopted into the poorer; also the younger the child at the time of adoption, the greater the gain in I.Q.

It is significant that the intelligence scores of adopted siblings reared in *different* foster homes correlated only .25, as compared with the *r* of .50 usually obtained between the scores of siblings raised together. Here it appears that the difference in environment has drastically reduced the resemblance. The correlation between foster-child

<sup>1</sup> The average of the intelligence scores made by father and mother.

I.Q. and foster-parent Otis S.A. score was .37; and the correlation between foster-child I.Q. and cultural rating of adopted home .48. These results indicate again a greater contribution of environment to differences in test performance than do the results of Burks. It should be said, however, that in Freeman's study selective factors in adoption may have operated to raise the correlation between the foster-parents' intelligence score and the child's I.Q., as well as the correlation between the rating of the foster-home and the child's I.Q. The children in Freeman's study were older at the time of adoption than those in Burks' study. More information was to be had, therefore, upon the achievements and probable capacities of the children to be adopted. It seems likely that the more intelligent and better educated foster-parents would avail themselves of any information upon the child and its antecedents, and hence tend to select the brighter children for adoption. The systematic operation of selective factors, however, is denied by Freeman.

The most clear-cut result growing out of both of these studies is the finding that the I.Q. does tend to rise after residence in foster-homes. In this respect the two studies show considerable consistency; Burks reports an average rise of five to six points under good environmental conditions, Freeman 7.5 points. However, these changes are, after all, relatively small, and indicate strongly that hereditary factors largely determine individual differences in performance upon the Stanford-Binet. Drastic changes in the environment will apparently operate to lower or to raise the I.Q. by as much as twenty points.

### THE "GENIUS" AND THE FEEBLE-MINDED CHILD

Before the advent of mental tests only striking examples of feeble-mindedness or genius attracted much notice. The large group of individuals not falling into either of these categories was simply classified as normal, or average. Such rough and superficial cataloging inevitably led to conceptions of normality, genius and feeble-mindedness as constituting "types," *i.e.*, groups possessing special characteristics which marked them off definitely (and qualitatively) from each other. Very little progress could be made toward the proper understanding of atypical individuals, particularly the feeble-minded, until their fundamental relationship to the average or normal individual was correctly understood. Mental tests may be said



to have demonstrated conclusively that the very bright and the very dull are not "types," but simply the extremes of a continuous scale of abilities.

### 1. The Superior Child

It is a popular notion that the superior child, although precocious in early years, will at maturity be no more intelligent—and often less so—than the average child. According to this view, the gifted child may learn more readily than the normal or slow child, but not more surely. Unfortunately for the parents of dull children, this "compensation theory" is not supported by evidence. Data collected by Baldwin (2) upon superior children indicate that the I.Q.'s of superior children remain fairly constant from year to year, the superior child tending to maintain his relative superiority to the average and the dull child, as he grows older.

By far the most extensive study of superior children by the mental test method has been conducted by Terman and his associates (50) at Stanford University. This investigation is so complete, and demonstrates so clearly the value of mental testing in research of this kind, that it will be quoted here in some detail. The main group in Terman's study consisted of 643 school children, all with Stanford-Binet I.Q.'s of 140 or above. As Terman points out, this group represents approximately the upper 1 per cent. of the elementary school population. Control groups of from 600 to 800 unselected school children of average age-grade location were selected for comparison with this superior group. All of the children—control and experimental—attended the same schools. The data collected on the gifted children included Stanford-Binet; N.I.T.; the Stanford Achievement Test; tests of general information and of knowledge about common games and sports; tests of interests; personality tests, as, for example, a revision of the Woodworth P.D. Sheet; objective tests of honesty and other character traits based upon the Voelker tests; tests of social attitudes; and finally, anthropometric and medical examination records. Each child was rated by its parents and teachers upon a variety of intellectual, emotional, social and moral characteristics. Information was also collected upon the cultural status of the home, education and occupation of parents, and the special achievements of relatives and ancestors.

Analyses of these extensive data indicated that in general the

superior child comes from a better home, has better educated parents and better educated relatives than the average child. The average cultural rating of the homes of the superior children upon the Whittier scale, for instance, was 22.94, as compared with 20.78 for the general population—and this in spite of the fact that the economic status of the superior group was by no means above the average of the localities from which the control group was drawn. The median school grade reached by the parents of the superior children was 12.2, and by the grandparents 8.9, in contrast to the norm of 6.9 reported for the Army draft as typical of the general population. The ratings of parental occupations (for intellectual requirements) on the Barr scale gave an average of 12.77 for the superior group as compared with 7.92 for the general population in the same districts.

The anthropometric and medical records, as well as the questionnaire data submitted by the children and their parents, indicated that the gifted children as a group are less sickly than the average, and are stronger in muscular development. Poor nutrition was reported in 7.2 per cent. of the control group and 2.6 per cent. of the gifted; nervousness in 16.1 per cent. of the control and 13.3 per cent. of the gifted. Indices of development showed the gifted group to be somewhat accelerated physically, as well as mentally. The gifted children were reported as having on the average begun to walk one month earlier, and to talk three and a half months earlier, than the average group.

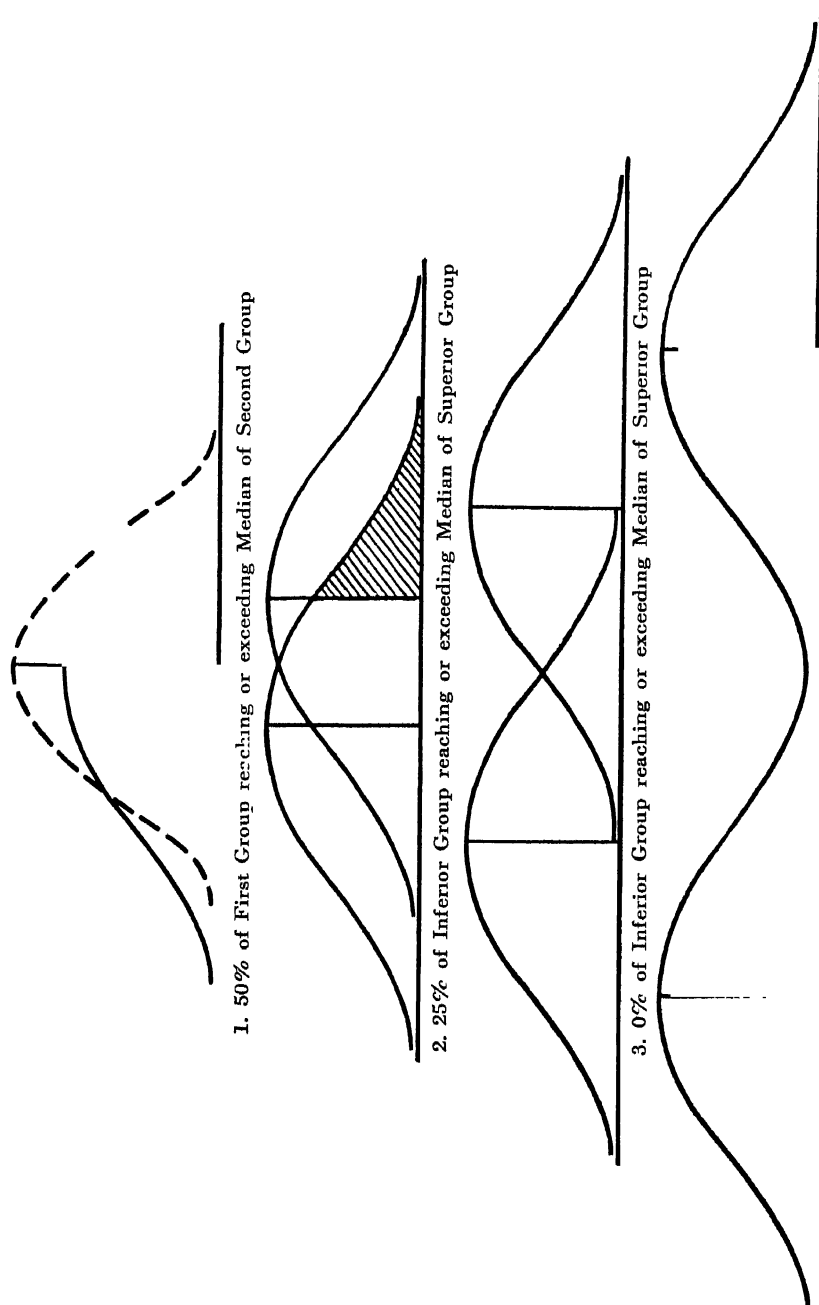
No evidence was found that the interests of the gifted child were more narrow or more specialized than those of the average child; analysis of their play life, for example, showed no marked deviations from the norm, except a tendency for the superior child to prefer games of children chronologically older than himself. The personality questionnaires and ratings indicated a distinct superiority of the gifted group. The gifted children were significantly higher than the control in the personality and character tests, in social attitudes and in trustworthiness.

Educationally, the gifted group was accelerated, 85 per cent. being ahead of their normal age-grade location, 15 per cent. normal, and none retarded. Scores on the Stanford Achievement Examination showed that many of the gifted children should have been in grades still higher than those in which they were placed. Although the gifted

group, on the average, did 40 per cent. better than the age norms on the Stanford Achievement Test, their actual grade location was only 14 per cent. better than the average. This retardation of bright children has been commented upon in Chapter I, (p. 30) and shows the need for special classes, and for more flexible promotion methods for the bright child.

## 2. The Feeble-minded

Mental tests have most often been employed with the feeble-minded in connection with problems of detection and placement; and for determining the type of training best suited to individual cases. Definite indications of the kind of training which should prove most fruitful to the mentally retarded or deficient may be found in studies of the relative retardation of the feeble-minded in different traits. All of the experimental evidence points to the conclusion that the dull child is not equally inferior in all respects, his deficiency being most marked in the abstract verbal functions, less so in simple mental processes requiring attention and discrimination, and least of all in physical and muscular development. In a pioneer study by Norsworthy (38), a series of sixteen mental, motor and sensory measurements were secured from 157 subjects, the majority of whom were between the ages of eight and sixteen. All of these children were inmates in institutions or members of special classes for the feeble-minded. Records were taken from several hundred normal children to be used for comparison. Norsworthy's results are expressed in terms of the percentage of the feeble-minded group which reached or exceeded the median of the normal group. In such a comparison, 50 per cent. "overlapping" indicates a complete correspondence within the two groups. (See Figure 15.) In four physical tests, *viz.*, height, weight, pulse rate and body temperature, the average percentage of the feeble-minded reaching or exceeding the median of the normal was 41 per cent.; in a test of equating weights it was 18 per cent.; in the "A" cancellation test, 9 per cent.; and in the a-t cancellation test, 1 per cent. Memory for unrelated words gave an overlapping of 6 per cent., and memory for related words 5 per cent. A selection from the Woodworth-Wells association tests, used as measures of higher or abstract abilities, gave the following percentages of overlapping: part-whole, 9 per cent.; genus-species, 9 per cent.; opposites, 0 per cent. Comparative studies of feeble-



4. No Overlapping of Inferior and Superior Groups  
Figure 15.—OVERLAPPING OF FREQUENCY DISTRIBUTIONS

mind and normal children in school subjects have given quite similar results (11, 35). The feeble-minded are most deficient in reading, composition and arithmetic; and least deficient in simple computation, penmanship, drawing, handwork and spelling.

To the psychologist, these results indicate definitely that the feeble-minded should be treated much like the normal, since they differ from the normal in *degree*, rather than in *kind*, of intelligence. A practical conclusion is that training for the feeble-minded and low-grade must place its emphasis upon training in manual skills and upon routine, rather than upon intellectual, activities in order to develop those capacities in which the feeble-minded has the best chance of success. The importance of training the feeble-minded in manual tasks, and in simple sensory discrimination, was emphasized as long ago as 1866 by the French physician Séguin (45), who demonstrated the significant progress that can be made by the low-grade in such activities. In modern institutions, all but the lowest cases of feeble-mindedness are kept busy at some productive task suited to their limited capacities. Besides the social usefulness of such work, the assignment of tasks which he can perform effects a favorable emotional attitude in the feeble-minded individual by arousing his self-confidence, interest in accomplishment and feeling of independence.

### GROUP DIFFERENCES

In no field of investigation have mental tests been more widely used than in the study of the differences among groups. Even before the present extensive development of, and interest in, intelligence testing, sensory, motor and even simple mental tests (56) were administered to individuals of both sexes and to various racial and national groups. If we add to these earlier studies the many recent investigations of group differences made with general intelligence tests and other psychological measures the total becomes exceedingly great.

The results of many studies of group differences in tested abilities are inconclusive or ambiguous, unfortunately, owing to the failure of the investigator to overcome difficulties which are nearly always present in such comparisons. The more important of these hazards may be listed as follows: (1) The unreliability and ambiguity of many tests; (2) the errors arising from the use of small and non-representative samples; (3) the failure to take account of overlap-

ping; and (4) wide environmental and cultural differences. These complicating factors will be illustrated briefly. The unreliability of many tests (p. 44) renders the differences between average scores non-significant statistically, while the meaning of the scores themselves is often not clear. To state, for example, that boys are superior to girls in "reasoning ability" means little, unless one knows whether the difference between the group averages is significant, and what the tests are from which this general conclusion was drawn. Even large differences between averages may arise from chance alone, while "reasoning ability" may refer to arithmetic problems or to logical problems of a syllogistic sort, or to some other very different test. When two samples are compared, they must not only be large enough to make comparisons valid, but they must also be truly representative of the populations from which they are drawn. A study by Brigham (6) illustrates the sampling hazard. Basing his conclusion upon the Army Alpha test scores made by foreign-born men enlisted in the United States Army, Brigham reported the intelligence of men born in South European countries (Mediterranean) to be less than that of men of Northern European ancestry (Nordics). The groups upon which these results were obtained are so small, in most cases not exceeding 400 or 500, that it is hard to conceive these groups to be representative of their respective nations. The factors determining immigration from Sweden and Italy, say, may differ so greatly that one may be getting the intellectually poorest from the one country and the average or superior from the other. One is not justified in concluding, therefore, from the test scores of 500 Swedes and 500 Italians that the Swedish nation is, in general, more intelligent than the Italian.

The failure to take account of overlapping is a potent source of error in comparing two or more groups. It may happen that the difference between the averages of a group of eight-year-old boys and a group of eight-year-old girls on a given test is five points in favor of the girls. But when one examines the distributions of scores, it usually appears that the differences *within* either group are far greater than the differences *between* average members of the two groups. Said differently, a high-scoring boy and a low-scoring boy will usually be much farther apart in ability than the average boy and the average girl. To draw the conclusion that one group is su-

perior to another, without taking into account the overlapping of the two groups, may greatly exaggerate, if it does not falsify, the facts.

We have already indicated on p. 23 how environmental differences, such as language, cultural status and training, affect the I.Q. In exactly the same way, such variations in background often preclude a fair comparison of two groups in other respects. Klineberg (31) has shown that the differences between the performances upon mental tests of American Indian and white children arise from fundamental differences in attitude toward the test, rather than from differences in ability; and the same is true of the differences in performance of white and of Oriental and Negro groups. Klineberg observes, in illustrating this point, that it is unreasonable to expect a Negro child to indicate as "True" the statement: "Silence is observed in libraries and churches" when all of his experience would indicate that the contrary is true. The upshot of this discussion is that comparisons of groups by means of tests, *e.g.*, boys and girls, Negroes and whites, are valid *only* when environmental factors which may affect test scores are equalized or made negligible. In the descriptions of the studies of group differences in the following paragraphs it must be remembered that results and conclusions are always to be evaluated in the light of the criticisms and limitations outlined above.

### 1. Sex Differences

Most investigators have found the average boy to be superior to the average girl in tests involving numerical or mathematical relationships, as well as in the ability to employ spatial and geometric concepts. Girls, in turn, are usually superior to boys in tests requiring memory and in the ability to employ language relationships quickly and precisely. To be sure, the sex difference found in a *single* study for any of these functions is often not significant statistically; but the consistency of the results reported by different investigators lends weight to the belief in the existence of a true difference.

Cameron (14) measured 500 boys and girls in fourteen secondary schools in England and Wales with various mathematical and spatial ability tests. The age range covered by the group was from thirteen years to sixteen years ten months. The tests included computation; arithmetic reasoning; mechanical processes in algebra; and the use of symbols in algebraic reasoning, as well as tests of geometric con-

struction and geometric reasoning. Cameron's results indicated a small but consistent difference in favor of the boys upon the whole battery. When boys and girls in *separate* schools were compared, the average percentage of the girls exceeding the boys' median was 39 per cent.; while in *co-educational* schools the average percentage of girls exceeding the boys' median was 40 per cent. The percentages of girls exceeding the median boy in the separate tests ranged from 29 per cent. (geometric space relations) to 59 per cent. (geometric deduction) when boys and girls in separate schools were compared. In the co-educational schools the percentages of girls exceeding the boys' median ranged from 26 per cent. (arithmetic problems) to 57 per cent. (mechanical algebraic processes). This study shows a small but consistent superiority of boys over girls in mathematical processes. The fact that the boys' superiority is as great in the co-educational schools (environment equal) as in separate schools (environment different) suggests that schooling, at least, is not the chief cause of the difference.

Terman (49) found that 15 per cent. more boys than girls passed the nine-year test of making change on the Stanford-Binet, and that 33 per cent. more boys than girls passed the fourteen-year test of arithmetic reasoning. Fauth (19), who administered a test of continuous addition to 1,214 German school children in Grades 1 to 8, reported that the girls did fewer problems in the given time and made more errors than the boys in every grade. Whipple (54), who tested a group of 834 high-school students with Army Alpha, found the boys to be significantly superior to the girls in the arithmetic test, the number series completion test, and the information test.

In performance tests, which seem to depend largely upon the ability to handle relations of a spatial sort (p. 69), boys again excel girls. Gaw (22) tested 100 school children, fifty-two boys and forty-eight girls, of average age thirteen and a half years, upon several of the Pintner-Paterson tests, the Porteus mazes, the Pintner Non-language Scale, the Stenquist Mechanical Aptitudes Test and the Kélley Construction Test. The Stanford-Binet was also administered to all of the members of the group. The I.Q. computed from the results of all of the performance tests gave a slight difference (three points) in favor of the boys; but analysis of the scores on the separate sub-tests revealed marked differences in favor of the boys in three of the Pintner-Paterson tests, *viz.*, the Picture Completion I, the



Triangle and the Diagonal tests. The girls excelled the boys in Cube Imitation, Substitution Learning, and Cube Construction—tests which depend, apparently, more upon memory and visual imagery than upon spatial relations. Much the same sex difference was found by Goodenough (24), who administered the Kuhlmann-Binet Scale and the Wallin Peg Boards to 300 children, ages two, three and four years. There were fifty boys and fifty girls at each age level. Although the girls were higher in Kuhlmann-Binet M.A., the boys were slightly ahead on the peg boards, their time scores being lower, but not always reliably so, than the girls' scores.

A consistent superiority of girls is shown in most studies of memory. Terman (49), in analyzing the Stanford-Binet results on kindergarten and school children, found over 10 per cent. more girls than boys passing each of the memory tests on this scale. Achilles (1), who tested adults and children in memory for words, for syllables, for proverbs and for geometric forms, reported female superiority in nearly every comparison made.

In studies of aptitude for, and facility in, language and language usage, most of the evidence indicates that girls are superior to boys. Trabue (53) reported that upon his Sentence Completion Test the girls' median scores were consistently higher than the boys' median scores for the same grade. On the Stanford-Binet, Terman found the average I.Q. of the girls to be higher than that of the boys up to age fourteen. Bonser (4), who tested 385 boys and 372 girls, ages nine to sixteen, upon a series of mathematical and verbal tests, found the girls superior in naming opposites, defining the meaning of words, and literary interpretation of passages of poetry. The boys excelled in the "mathematical judgment," and in the "best reasons" tests. L. W. Pressey (42) has reported results obtained upon the Pressey Group Intelligence Scale for 880 school children, ages nine to fourteen years. The average score of the girls upon the whole test was higher than that of the boys at each age. The girls were superior in rote memory for words, naming opposites, analogies, word completion, dissected sentences and moral classification—the last test probably measuring reading comprehension, primarily. The boys excelled in the arithmetic test at all ages, and in the information test from age eleven to sixteen. Book and Meadows (5) have reported results obtained with the Pressey Group Intelligence Scale upon 2,422 boys

and 3,503 girls, high-school seniors, ages sixteen to twenty-three. The boys were superior in arithmetic and information; the girls slightly superior in word completion, dissected sentences and logical memory. The relative performance of the two sexes on the sub-tests agrees in general with the study by Pressey described above. But in the high-school group the boys scored higher on the whole scale, whereas in the elementary school group the girls scored higher. Two factors may be mentioned briefly which singly or together might lead to this result. In the first place, the high-school boys probably represent a more highly selected group than the girls, since dull boys tend to drop out of school earlier than dull girls; the fact that there were more girls than boys in the high-school group supports this hypothesis. Again, girls may develop more rapidly than boys and hence reach intellectual maturity earlier. If this be the case, the younger girls in Pressey's study would do better than the boys, because of developmental acceleration; but this advantage would tend to disappear or to be reversed in the older groups. There is considerable evidence that girls mature earlier than boys in physical traits (3). And the presumption is that this is also true of mental development, although the evidence on this point is not conclusive.

One of the most striking sex differences is that reported by Brigham and Brolyer (7) from the results secured with the scholastic aptitude tests. These tests were administered in 1930 to 4,214 boys and 3,363 girls, all candidates for entrance into various colleges. Upon the verbal tests of the battery, the difference in favor of the girls was 11.34 times the S.D. (diff.), a highly reliable difference, statistically. A difference equally as large in favor of the boys was found on the numerical tests of the battery, the difference here being 15.27 times the S.D. (diff.)

The trend of the differences reported in this section bears out our earlier statement that, in general, girls do better on language and memory tests, and boys do better on mathematical, spatial and information tests. These results do not necessarily imply any innate or hereditary difference in constitution; but they do reflect an interesting difference under existing environmental conditions. Part of the sex difference found may be hereditary, but it is probable that a large—if not a major—part is to be attributed to differences in training, interests and outlook.

## 2. Race Differences

(a) *The American Negro*: Comparisons of Negro and white have been especially numerous in the United States owing to the fact that the Negro has lived in this country for 300 years; speaks English as his native tongue; and, in a broad sense, is subject to the same environmental influences as the white. In 1915, Ferguson (20), in a pioneer study, tested 421 Negro school children and 486 white school children in three cities in Virginia. The tests employed were the Woodworth-Wells Analogies, or Mixed Relations Test, the Trabue Language Completion Test, the "A" Cancellation Test, and a stylus maze. In analogies and sentence completion the Negro children were markedly inferior to the white; but in cancellation the differences were inconsistent, the Negro girls excelling the white girls, and the Negro boys ranking about the same as the white boys. In the stylus maze the Negro children were slightly more accurate, but somewhat slower than the white. When they worked as rapidly as the white children, however, they made the same number of errors. When the Negroes were classified on the basis of skin color, in order to show the amount of interbreeding with the whites, a positive relationship was obtained between all of the scores on the tests and the apparent amount of white blood. Ferguson's conclusion was that in general the performance of Negroes is about three-quarters as efficient as that of whites with the same amount of training. A contrary result has been reported when performance tests were used in which language-training and educational advantages play a minor rôle. Klineberg (31) administered four of the Pintner-Paterson Performance Tests<sup>1</sup> to 200 Negro and 100 white boys in New York City, ages eleven to sixteen, and 139 Negro and twenty-five white boys, ages seven to sixteen, in a rural district in West Virginia. His results showed the Negroes to be equal or slightly superior in accuracy to the whites, but slower in their performance. No relationship was obtained between test scores and amount of white blood, as judged from negroid characteristics, such as skin color, lip thickness, nose breadth, etc. The New York City Negro children were faster than the West Virginia Negroes, a close relationship being found between length of residence in New York City and speed upon the test. This result the author takes as indicating the speeding effect of an urban environment upon test performances of this sort.

<sup>1</sup> These were Mare and Foal, Casuist, Triangle and Healy "A." See p. 77.

One of the most extensive studies of the comparative abilities of Negro and white is that of Peterson (40), who tested 1,726 whites and 1,424 Negro school children in three southern states. Peterson used a series of group general intelligence tests, including the Pressey Mental Survey Scales, the Otis Group Intelligence Scale, the Haggerty Intelligence Examination Delta I and the Myers Mental Measure. The average Negro I.Q. on the group tests was .68 as compared with .88 for the whites. Comparisons of separate ages and grades indicated consistent Negro inferiority upon all of the tests. On the Peterson Rational Learning Test, which is relatively independent of training and language ability, a significant difference appeared in favor of the white children. This result was obtained in comparing 299 white and 314 Negro school children.

Pressey and Teter (43) administered the Pressey Group Test to 187 Negro school children in Grades 3 to 12. A comparison of the total scores of the Negro children with the white norms showed an average retardation of the Negroes of about two years. The scores on the separate sub-tests placed the Negro consistently below the white child. Negro inferiority was least in tests of rote and logical memory, arithmetic problems and practical judgment; and greatest in tests requiring literary ability, word knowledge and information. It is probable that lack of comparable environmental opportunities, difficult access to libraries and books, will explain part, at least, of the Negro inferiority in these tests.

The average score of native-born whites upon Army Alpha was fifty-nine (57), that of northern Negroes thirty-nine, and that of southern Negroes, twelve. For the illiterates in these three groups, the average score on Beta was forty-three for the whites, thirty-three for the northern Negroes and twenty for the southern Negroes. The relative standing of the three groups is the same on Beta as on Alpha, but the differences in average score are smaller. This checks up with our earlier finding that the Negro is much less inferior (if inferior at all) to the white in performance tests, than in "verbal" general intelligence tests. The superiority of the northern over the southern Negro in all of the tests has been variously explained (41). It is doubtful whether selective migration of the more intelligent and ambitious Negroes (which has not been proved) can account for all of the difference. Better education, as well as better social and occupational opportunities, seems to be a more reasonable explanation.

In the special field of music, Davenport and Steggerda (18) report the Negro to be superior to the white in tests of pitch discrimination, sense of rhythm and tone memory. Their subjects included ninety adults and 300 children.

By way of summary, it may be said that Negroes tested in the United States are generally inferior to the whites in verbal tests of general intelligence. The Negro is most inferior in tests demanding abstract reasoning and language knowledge and usage; he is equal, and sometimes superior, to the white in tests of memory and concrete practical judgment, and in accuracy upon performance tests. In simple tests of musical capacity the Negro is apparently superior to the white. Amount of white blood, as indicated by skin color and the absence of negroid characteristics, is apparently uncorrelated with performance upon intelligence tests, though the evidence upon this point is not conclusive. Whether the inferiority shown by the Negro upon mental tests is actually a matter of poorer native equipment rather than the result of more meager environmental opportunity, is still an unsettled question.

(b) *The American Indian*: Comparative studies of the abilities of the American Indian and the native white are rendered extremely difficult by wide differences in training, culture and attitude toward life. In those studies wherein environmental factors were at least roughly equated, the Indian appears to test somewhat below the white. Hunter and Sommermeier (28) administered the Otis Group Intelligence Test to 715 Indians (including the representatives of sixty-five different tribes) attending the Haskell Indian Institute in Kansas. Records of immediate ancestry were available for 711 members of the group, from which the degree of inter-mixture with the white race could be approximately determined. The median score of the Indian group on the Otis test was 82.64, as compared with the

TABLE IX  
COMPARATIVE SCORES ON THE OTIS GENERAL INTELLIGENCE TEST MADE BY  
INDIANS OF DIFFERENT DEGREES OF WHITE INTERMIXTURE

(from Hunter and Sommermeier (28))

Percentile	$\frac{1}{4}$ Indian Blood	$\frac{1}{2}$ Indian Blood	$\frac{3}{4}$ Indian Blood	Pure Indian Blood
25 .....	77.25	68.00	56.31	35.80
50.....	109.30	91.47	77.75	67.46
75.....	127.90	117.90	108.30	94.35
No. of Cases .....	112	192	142	265

white norm of 122.58. Pure-blood Indians ranked lowest, the test score increasing regularly with increase in proportion of white blood. (See Table IX.) A correlation of .51 was obtained between the Otis test score and the percentage of white blood when variability arising from age and school grade was held constant. That improved social surroundings influence the Indians' Otis scores is indicated by the positive correlations (.28 to .42) between total score and number of months in school. An analysis of the various sub-tests in the Otis scale showed the Indian inferiority to be greatest upon the highly verbal tests (those most susceptible to training) such as opposites, analogies, matching proverbs and narrative completion; and least upon tests of memory, of following directions and tests involving geometric or spatial relations.

Jamieson and Sandiford (29) examined 717 Indian school children in Ontario, Canada, using the following tests: N.I.T., Pintner Non-language Scale, Pintner-Paterson Performance Scale, Pintner-Cunningham Intelligence Test, and the Stanford Achievement Test. All of the children could speak English, although their ability was reported as somewhat below that of the average American child. Children in whose homes English was spoken exclusively were classified as the "English-speaking group," and their performance on the various tests was compared with that of the other group. The social status of the Indian group as a whole was considerably below that of the white, the average rating of the Indian homes on the Chapman Socio-economic Scale being 13, as compared with the white norm of 56. Upon the Stanford Achievement Test, the Indian children were decidedly inferior to the white norm, a part of which inferiority may be directly attributed to language difficulty, poor study conditions and irregular school attendance. Upon the Pintner-Paterson Performance and the Pintner Non-language Scales the Indian children were nearest to the white norms, their median I.Q.'s on these tests being 96 and 97, respectively. The median Indian I.Q. on the Pintner-Cunningham test was .78, and on the N.I.T., .80. Upon all of the tests except Pintner-Paterson, the "English-speaking" Indian children surpassed the other Indian children. These comparisons of Indian and white upon verbal and performance tests indicate clearly the large influence of the language factor, as well as the effect of wide differences in environmental and cultural status.

Klineberg (31) administered six<sup>1</sup> of the Pintner-Paterson tests to 120 Indian children upon the Yakima Reservation, and to a control group of 107 white school children, ages seven to sixteen, living on the Reservation. The same tests were also given to 136 Haskell Institute Indians, ages seven to twenty-one. On all of these tests the Indians worked more slowly than the whites; but they worked more accurately at each age level. The Haskell Indian group worked faster than the Yakima Indians. This last result, and the lack of correlation between speed on the tests and white intermixture, led Klineberg to suggest that the differences in performance probably resulted from environmental factors rather than from native differences. The author points out that speed does not have for the Indians the same significance and importance which it possesses for the whites.

The results from mental tests given to the American Indian are, in general, quite similar to those results obtained with Negroes. The Indian is consistently inferior to the white in language tests; considerably less inferior, or even superior, upon performance tests. Language difficulties, inferior schooling, differences in attitude, in interests, in temperament and in incentives loom so large, that an evaluation of the part played in Indian-white mental differences by immediate ancestry is well nigh impossible. The negligible differences between Indians and whites upon performance tests, which are largely independent of environment, suggest that the influence of hereditary differences would be very slight, if the environments were really comparable.

(c) *Oriental Groups in America*: There have been a number of studies of Oriental groups residing within the United States. Darsie (17) carried out an extensive investigation upon American-born Japanese children living in California. His main group, all of whom lived in cities and reported English as the language most familiar to them, totaled 658 children between the ages of ten and fifteen. Tests administered were the Stanford-Binet, the Army Beta and the Stanford Achievement. Teachers' ratings were also secured upon nineteen social, personality and other traits. Upon the Army Beta, no significant differences were found between the white and Japanese children up to the age of twelve. Beyond twelve years, the Japanese

<sup>1</sup> Mare and Foal, Casuist Form Board, Triangle, Healy Puzzle "A," Ship, and Knox Cube Tests (p. 77).

were somewhat superior. Upon the separate sub-tests of Beta, the Japanese were superior in digit-symbol learning and number comparisons, but were significantly inferior in picture completion. The Japanese children had a mean I.Q. of 89.5 upon the Stanford-Binet as compared with 99.5 for the white children from the same district. This difference, however, was found to be attributable to a few highly "verbal" tests, which probably favored the white children. The Japanese children, for instance, were reliably superior to the white children on the Stanford-Binet tests which involve sustained attention and visual perception, *viz.*, induction, paper cutting, inclosed boxes and code writing. In an effort to decide the degree to which success in each of the Stanford-Binet tests depends upon language comprehension, Darsie had seven psychologists rate each of the tests for "verbalness." Each test was then given a rank based upon the average judgment of these seven experts. The coefficient of correlation between the ranking of the tests for language comprehension and for degree of Japanese inferiority was .87, which shows that the Japanese tended to be most inferior on those tests judged by experts to be most verbal. As shown by their Stanford Achievement Test scores, the Japanese children were about six months retarded. The only school subjects in which these children were normal were arithmetic computation and spelling, in which, probably, their tendency to give close attention and their keen perception offset other disadvantages. It is interesting to note that Japanese children were rated by their teachers as superior to the "standard American child" in modesty, emotional stability and esthetic appreciation; and inferior in self-confidence, command of language and intellectual capacity—this last rating reflecting, probably, their somewhat inferior school attainments.

Sandiford and Kerr (44) gave the Pintner-Paterson Performance Scale to 500 Chinese and Japanese children in the Public Schools of Vancouver, B. C. The average I.Q. of the Japanese boys was 115.4, and of the Japanese girls 112.8; while the average I.Q. of the Chinese boys was 107.8, and of the Chinese girls 107. These results indicate that, on tests which make little demand upon language comprehension, the Chinese and the Japanese school children do as well as, or better than, comparable white children. As noted above, the Japanese appeared to be especially superior upon tests involving quick attention and perception.



The equality and frequent superiority of Orientals in mental tests has been usually attributed to selective migration, the assumption being that only the more progressive and more intelligent individuals leave Oriental countries to come to the United States. This may have been true in times past, but it is probably not true today—and may not have been true then. As already stated, factors determining emigration vary from country to country, some countries sending their best, and others their worst elements. This considerably complicates comparisons of foreign-born groups in the United States.

(d) *Natio-racial Differences*: In attacking the problem of racial differences, psychologists have often failed to distinguish between racial divisions and national divisions. As commonly used, the term race, as for example the Negro race, refers to a group of people who possess common inherited biological and anatomical characteristics, such as head shape, hair, eye and skin color, and other facial and bodily similarities. A nation, on the other hand, is a geographical or political group, e.g., French or Japanese, living within the boundaries of a common country and having a common government. Obviously, a nation may be a mixture of many diverse racial elements.

Modern anthropologists divide the white peoples of Europe into three great racial sub-groups: Nordic, Alpine and Mediterranean. The Nordic is characterized by blondness (blue eyes, light hair), tall stature, and dolichocephaly (long-headedness);<sup>1</sup> the Mediterranean by brunetness (dark eyes and hair), short stature, and dolichocephaly. The Alpine falls in between these two extremes, being intermediate in eye and hair color and medium in height. In head shape he is brachycephalic<sup>1</sup> (broad-headed). Every European nation includes in its racial composition all three of these strains. In order to study racial differences as between European nations, therefore, individuals should be classified according to those physical resemblances which are indicative of common heredity, rather than according to arbitrary political groupings. Two studies of racial differences within the white race will be described in this section.

Hirsch (27) gave mental tests to 5,504 school children in the public schools of Massachusetts. These children were in Grades 1 to 9, and ranged in age from five and one-half to eighteen. While the children were all American-born, the countries of birth designated by their parents represented a total of fifteen nations. All of the

<sup>1</sup> See Chapter I.

children tested attended the same schools as the children of American-born parentage, and were quite well equated with them in social status and parental occupation. The first-grade children were given the Pintner-Cunningham General Intelligence Test, the second and third grades the Dearborn Group Test of Intelligence, A, and the fourth grade and up the Dearborn Group Test C. Extended time limits were allowed in order to eliminate the influence of speed as much as possible.

The children were first classified according to the *national origin of their parents*. When these groups were compared, the differences between the average scores were nearly always statistically reliable. The children were next classified roughly into the three racial sub-groups, Nordic, Alpine and Mediterranean, irrespective of the national origin of their parents. Eye and hair color were used as the criteria for classification, the three final groupings being: (1) "blond," i.e., those with light hair and blue, gray or hazel eyes; (2) "brunet," those with black hair and gray, hazel, brown or black eyes; and (3) "mixed," those children with brown hair, and any eye-hair combination not included in the first two groupings. Group I was taken to represent the Nordic racial strain; Group II the Mediterranean, and Group III the Alpine, and any racial mixtures of the other two groups. When these three racial sub-groups, as defined above, were compared upon the intelligence tests, it appeared that race made far less difference than nationality, so far as I.Q. differences were concerned. The range of differences in average I.Q. for the three racial sub-groups *within a single nationality*, for example, was very small, the greatest range being 6.7 (in the Polish group) and the smallest .6 (in the French-Canadian group). But the I.Q. range of the different nationalities *included within each racial type* was extremely large, being 14.8 within the blond group, 21.3 within the mixed, and 25.6 within the brunet. Hirsch's classification of his subjects into "racial" groups is admittedly open to criticism. If it may be accepted, however, as tentative, his results suggest strongly that scores on mental tests are governed to a much greater degree by those differences in environment which characterize the different national groups, than by those ostensible hereditary differences which characterize racial groups.

In a more recent and better controlled study of natio-racial differences, Klineberg (32) gave six of the Pintner-Paterson Per-

formance Tests to ten groups of children, each group consisting of 100 boys, ten to twelve years old. These tests were the Triangle Test, the Healy "A," the Two Figure board, the Five Figure board, the Casuist board, and the Knox Cube Test. The experimental groups included city children drawn from Paris, Hamburg and Rome, and country children from the rural districts of France, Germany and Italy. Only those children who exhibited the distinct physical make-up characteristic of one of the three racial types, Nordic, Alpine or Mediterranean, were included in the rural groups. The criteria for classification into racial groups were eye color, hair color and cephalic index (Chapter I). The average score (point scale) on the six Pintner-Paterson Tests as well as the range of scores for each of the ten groups is given in Table X.

TABLE X  
COMPARATIVE TOTAL SCORES UPON THE PINTNER-PATERSON SCALE OF PERFORMANCE TESTS OF TEN GROUPS  
(from Klineberg [32])

Group	N	Average	Range	S D.
Paris (city) . . . . .	100	219 0	100-302	46 2
Hamburg (city) . . . . .	100	216 4	105-322	45 6
Rome (city) . . . . .	100	211 8	109-313	42 6
German Nordic . . . . .	100	198 2	69-289	49 0
French Mediterranean . . .	100	197 4	71-271	45 6
German Alpine . . . . .	100	193 6	80-211	48 0
Italian Alpine . . . . .	100	188 8	69-306	48 4
French Alpine . . . . .	100	180 2	72-296	46 6
French Nordic . . . . .	100	178 8	63-314	56 4
Italian Mediterranean . . . .	100	173 0	69-308	54 2

The test data show, in the first place, that differences among the three city groups are small and unreliable, overlapping being exceedingly great. In the case of the rural children, if we compare the three racial groups irrespective of nationality, small and unreliable differences appear. The difference divided by the S.D. (diff.) is .63, as between the Nordic and Mediterranean; .20 as between Nordic and Alpine, and .51 as between Alpine and Mediterranean. But if national groups are compared, irrespective of racial composition, the differences are far greater. The difference divided by its S.D. was 2.97 as between German and Italian children; 2.25 as between German and French children; and 1.03 as between French and Italian. The most surprising differences were found between those of the *same* race but of *different* nationality, the difference between the scores of the French Mediterraneans, and the Italian

Mediterraneans, for example, being 3.44 times the S.D. (diff.). Perhaps the most striking difference to appear between *any* two groups was obtained by comparing *all* of the city children with *all* of the country children, irrespective of race or nationality. The average score of the city children was 215.7 points on the performance scale, and that of the country children 187.1, the difference being over eight times as large as its S.D. (diff.).

The results of these two studies indicate strongly, as we have pointed out above, the powerful effects of environment, *i.e.*, training, education, religion, culture, ideals and outlook upon life. When individuals are classified upon the basis of environmental similarity, *i.e.*, as living in the city, or as living in the country, or as living within the boundaries of a single nation, the differences among such groups in mental test performances are much greater than when individuals are classified according to common heredity, or common ancestry inferred from similarities in physical traits.

#### MENTAL TESTS IN VOCATIONAL SELECTION AND VOCATIONAL GUIDANCE

A few typical tests used in special fields have been described in Chapter IV. These tests, as well as tests of general intelligence and especially designed batteries or trade tests, have been widely used in vocational selection and guidance. In this section several studies illustrative of these procedures will be described.

On p. 37, the intelligence ratings of various occupational groups, as determined from the Army Alpha Test, were presented. Two main facts stand out clearly in this table. In the first place, the amount of overlapping in Alpha intelligence scores from one occupation to another is very great. This indicates that an individual cannot be classified on the basis of Alpha score much more precisely than as falling within the professional, clerical, skilled labor or unskilled labor groups. Any finer classification would be extremely dubious. In the second place, clerical and business groups, accountants and bookkeepers, for example, have higher average general intelligence test scores than have workers in the skilled trades, such as electricians and machinists. Such differences in "intelligence" are in part, at least, owing to the fact that the first group is more accustomed in its every-day work to the kind of activities called for by the Alpha Test. A paper-and-pencil test of verbal in-

telligence may be an adequate means of differentiating good clerks from bad clerks, but an extremely poor way of separating good mechanics from bad ones. In vocational selection, the kind of ability required, whether abstract, social or mechanical (p. 4), must always be considered in selecting a test.

The general intelligence test when supplemented by other evidence has proved to be extremely useful in the vocational placement of individuals who are intellectually sub-normal, but not sufficiently low-grade to warrant commitment to an institution. Burr (10), in 1924, reported data upon 375 girls, ages fifteen and one-half to twenty-two, who had been examined with the Stanford-Binet, and various trade tests at the Vocational Adjustment Bureau for Girls in New York City. Mental ages ranged from six to seventeen years. After being examined, the girls were placed in different kinds of industrial work by the Bureau, and their later success at their work checked against their test records. The criterion of success on a job was set as the ability to hold the job for at least three months. Minimum mental age requirements were then determined for a number of specific industrial tasks. In the simple operation of packing, for example, a mental age of seven years and six months was found to be adequate, if the task involved the packing of small articles not damaged by careless handling, such as the placing of powder-puffs in oiled envelopes. The packing of objects which required careful handling, such as hairnets, demanded a minimum mental age of nine years and nine months; and when in addition to packing, stock-keeping, labeling and checking were required, ten years and five months was the minimum mental age.

In machine-operating jobs, such as sewing by machine, it was found that a mental age of at least thirteen years was required if accidents to the operator, the material or the machine were to be avoided. Of course, the mere possession of the minimum mental age requirement does not insure success in a given occupation. Other factors such as emotional control, physical health and stability of interests must be considered. However, the establishment of a minimum mental age requirement does permit the immediate weeding out of those individuals who would almost certainly fail in the given vocation.

Another study in which minimum mental age requirements for different occupations proved to be useful is that of Cowdery (16).

This investigator administered the Stanford-Binet to 578 boys, ages twelve to nineteen years. All were inmates of a reform school, and were being taught trades by apprenticeship. Cowdery first determined the minimum mental age requirements for each type of work. He then investigated the relationship between Stanford-Binet mental age and success in learning a job, in groups of boys all of whom reached or surpassed the minimum I.Q. set for that particular task. The trades studied fell into three groups with respect to the relationship between I.Q. and vocational success. In the first group were jobs which required a certain amount of skill and responsibility, such as office boy, hospital assistant, tailor's apprentice, carpenter's apprentice, and the like. The correlations between I.Q. and success in these jobs were positive and often significant, varying from .98 to .23. In the second group, the "neutral" group, were such jobs as dairy, kitchen, bakeshop and dining-room work, tending flower garden and acting as teamster. There was no appreciable relationship ( $r$ 's ranged from .09 to  $-.15$ ) between I.Q. and success in these jobs. The third group included work in the laundry, simple work in the garden and other tasks done under strict supervision. Success in these routine tasks was *negatively* correlated ( $r$ 's ranged from  $-.23$  to  $-.38$ ) with I.Q., the best work being done by those possessing meager "verbal" intelligence. It should be added, again, that in no group was the degree of success wholly dependent upon the general intelligence test rating.

In the studies so far described in this section, vocational selection or vocational guidance was made on the basis of general intelligence tests. We shall now describe a study in which vocational tests, especially constructed for the purpose, were employed. One of the most successful applications of a vocational test battery to a specific problem is illustrated in the investigation of Snow (47). Snow observed and tested nearly 3,000 drivers and 1,000 new applicants for the Yellow Cab Company of Chicago. First, a thorough analysis was made of the qualifications and traits apparently necessary for successful taxi driving. Snow then selected or devised tests which were intended to measure carelessness, recklessness, emotional stability, judgment and physical and sensory defects. This test battery was given to 311 applicants; and on the basis of the scores made upon only two of the tests (emotional stability and an intelligence or judgment test) thirty-four applicants were classified as probably

unsatisfactory, the rest being marked probably satisfactory. When these 311 men were followed over a period of ten weeks of actual driving, it was found that 64 per cent. of the unsatisfactory group had accidents in which the driver was clearly to blame, while only 33 per cent. of the satisfactory group had such mishaps. Those having more than two accidents constituted 38 per cent. of the unsatisfactory, and 12 per cent. of the satisfactory group. In spite of a steady increase in mileage covered, Snow was able to show a decided decrease in accidents for the year 1925, when unfit applicants were eliminated by means of psychological tests.

### MENTAL TESTS UPON DELINQUENTS AND CRIMINALS

In studying delinquents and criminals, the main psychological task has been the search for traits or characteristics which would serve to mark off such persons from the non-criminal population. One of the earliest and best known theories of criminality is that of Lombroso (34) who advanced the view that the criminal differs physically and mentally from the rest of mankind. According to Lombroso's theory, the criminal represents a reversion to a more primitive evolutionary level. Cases were brought forward and analogies cited to show the physical resemblances of criminals to primitive peoples and to lower animals; and various "stigmata of degeneracy" were enumerated which were supposed to be characteristic of the physical make-up of the criminal.

Lombroso's conception of criminality, like other attempts to catalog or pigeon-hole individuals into "types," was overthrown by later experimental investigation. Largely instrumental in this was the work of Goring (25) who collected physical data, as well as estimates of intelligence, upon 3,000 English convicts. Except for the fact that the convicts were—on the average—from one to two inches shorter than the normal, and from three to seven and one-half pounds lighter, no consistent differences in physical traits appeared. Estimates of feeble-mindedness were very high, however, 10 to 20 per cent. of the criminal group being adjudged feeble-minded, as compared with an estimate of  $1\frac{1}{2}$  per cent. feeble-minded in the general population. Goring concluded from his data that intellectual inferiority is the chief cause of crime, and this conclusion has been approved by many writers on the subject (23) and is still widely held. There are several considerations, however, which make this

view appear to be an oversimplification of the problem, and which tend to disprove the view that crime is simply the result of low intelligence (12, 26). In the first place, estimates or judgments of feeble-mindedness are open to considerable error; and when judges or legal officers are asked to estimate the intelligence of those brought before them, this error will probably be intensified in the direction of underestimation. It has frequently been pointed out that in considering only those actually committed to penal institutions, we are not getting a representative sampling of the criminal population, since the more intelligent law-breakers are less likely to be apprehended. Again, when general intelligence or other tests have been employed as a means of comparison, the criminal and the non-criminal groups have rarely been equated for environmental influences, making conclusions as to differences in intelligence extremely doubtful.

Criminals must be compared with non-criminals drawn from the same social and economic classes if fair conclusions as to differences in intellectual ability are to be reached. This is well illustrated in a study by Bronner (8), who gave a series of mental tests, *e.g.*, easy opposites, memory for hard words and passages, and completion, to four groups of girls, *viz.*, delinquents, servant girls, evening-school students and college students. The percentage of delinquent girls who reached or exceeded the median of the college group in five tests was 2.7; the percentage reaching or exceeding the evening-school median, 22.3; and the percentage reaching or exceeding the servant girls' median 56.7. The delinquent girls were clearly inferior to the college and evening-school girls in all the tests, but when compared with a non-delinquent group of approximately the same social status as themselves, they were clearly not inferior.

In the following section several representative studies of adult criminals and juvenile delinquents will be presented. It would be impossible in a short space even to sample the mass of psychological literature on this topic. Hence, since our primary purpose is to illustrate the value of psychological tests, we shall confine our discussion to the rôle of mental tests in studies of this kind.

### 1. The Adult Criminal

One of the most exhaustive surveys of intelligence test records upon large groups of criminals is the study of Murchison (37), who



gave Army Alpha to the inmates of penitentiaries in five states. These states were Illinois, Ohio, Indiana, New Jersey and Maryland, and the subjects were 3,954 white, native-born, male criminals. In addition to the study of this group, separate surveys were made of foreign-born, Negro and women criminals. The general intelligence scores made by the criminals were compared with the Army norms for the same state. Data were also collected upon educational and occupational status, type of crime, number of crimes, age, physical condition, marital status, religion, length of incarceration, *etc.*

No consistent differences appeared when the average intelligence test scores of criminals were compared with the Army norms for the same states. Intelligence, however, did seem to be related to the type of crime. Murchison, for example, classified the crimes into seven major groups, and computed the percentage of Army Alpha scores falling in each group, which could be classed as superior, *i.e.*, *above* grade C of the Army norm, and the percentage classed as inferior, *i.e.*, falling *below* grade C. Percentages of criminals classified as of superior or inferior intelligence upon this criterion are presented in Table XI for each type of crime. It will be seen that

TABLE XI  
PERCENTAGES OF CRIMINALS COMMITTED FOR VARIOUS OFFENSES  
WHO WERE OF SUPERIOR AND INFERIOR INTELLIGENCE AS JUDGED  
BY THE ARMY ALPHA EXAMINATION  
(from Murchison [37])

Type	Superior	Inferior
Fraud . . . . .	52 9	22 0
Force . . . . .	40 5	30 6
Thievery . . . . .	40 7	31.8
Statutory . . . . .	31 7	31.0
Physical Injury . . . . .	35 0	36 9
Dereliction . . . . .	35 3	43 1
Sex . . . . .	26 3	47.6

the group committed for fraud ranks highest in general intelligence score; while those committed for crimes of a social nature or for sex crimes rank lowest. Recidivists (repeated offenders) scored higher than first offenders. The modal chronological age of the whole group of criminals fell between twenty-one and twenty-five years, nearly one-half being twenty-five years old or younger. In amount of schooling the criminal group fell far below the Army norms for the population in general. The percentage of criminals who had had more than an elementary school education was 16.2 per cent. as compared with the 27.0 per cent. of the Army draft; the

percentage having less than an elementary school education was 47.3 per cent. as compared with 28.2 per cent. for the Army draft.

Results from the Alpha intelligence tests given to Negro criminals were in general fairly similar to those for the white. The Negro criminals in a given state were only slightly inferior to the Negro draft from that state. The foreign-born criminals, also, were not very inferior in Alpha score to the foreign-born Army draft. It is difficult to interpret the data obtained from women criminals, since there are no adequate Alpha norms for women. In addition, the group of criminal women studied was small, consisting of only eighty-five white and forty-one Negro women. The most clear-cut conclusions which we may draw from the small section of Murchison's data which we have considered are (1) that criminals are less well educated but probably not inferior intellectually to the general population; (2) that the type of crime committed is related to the intelligence level of the offender.

## 2. Juvenile Delinquency

Slawson (46), in 1926, tested 1,543 delinquent boys, average age fifteen years, five months, in four institutions in New York State upon the N.I.T., the Thorndike Non-Verbal Test, the Stenquist Mechanical Aptitudes Test, and the Woodworth-Mathews Personal Data Sheet. Data were secured also on height, weight, sensory defects, physical strength and vital capacity. Slawson treated his results in two ways: first, the scores made by the delinquent boys were compared with the norms reported by the author of each test; and, secondly, the scores on each test were correlated with estimates of "degree of delinquency," based upon number of arrests and the severity of the penalty for each offense. Correlations between estimates of delinquency and test scores were too low to indicate any real relationship. In verbal intelligence, *i.e.*, upon the N.I.T., the delinquent boys were inferior to the test norms, only 18 per cent. reaching or exceeding the normal age medians. Slawson points out, however, that this apparent inferiority may have resulted from the fact that the groups upon whom the N.I.T. norms were established were hardly comparable in social status to his delinquent boys. Evidence favoring this explanation was found in comparisons made within the delinquent group, which indicated that race, nationality and paternal occupation made for large differences in N.I.T. score. In addition,

one group of delinquent boys, who had been tested upon Stanford-Binet, was compared with a dependent but non-delinquent group of boys in an institution for dependents. This latter group was quite comparable to the delinquent boys in social status. In this comparison, the delinquent boys proved to be superior in general intelligence scores to the non-delinquent. This suggests that much of the apparent inferiority of the delinquents upon intelligence tests may be the effect of their poorer social and environmental status.

Upon the Thorndike Non-Verbal Test, 32.6 per cent. of the delinquent boys reached or exceeded the test norm; and upon the Stenquist Mechanical Aptitude Test 49 per cent. in one institution, and 44 per cent. in another, reached or exceeded the test norms. It is evident that the more the test depends upon concrete, mechanical ability and the less upon verbal or scholastic aptitude, the better the performance of delinquent boys relative to the standardization groups. Concrete and manipulative tasks are those in which social status, education and training count least; and it is in these tasks that the delinquent boys are the least handicapped.

As judged by the emotional stability questionnaire, the delinquent group exhibited more neurotic traits than do unselected school children. The percentage of delinquent boys reaching or exceeding the normal mean was 84.4 per cent.; that is, 84.4 per cent. of the delinquents gave a larger number of neurotic responses than the average non-delinquent child of the same age. Differences in nationality as well as in social status, as between the delinquent and non-delinquent groups, probably affected these results somewhat, since children of foreign parentage tend to give more neurotic answers to the Woodworth-Mathews questionnaire than do native-born children. But when the delinquents of American parentage were compared with Mathews' non-delinquents (among whom were some children of foreign parentage) the delinquent group still had an average score of sixteen neurotic answers as compared with an average of nine for the non-delinquents. Similarly, in a group of Hebrew Orphan Asylum children (all non-delinquents) tested by Slawson, only 55 per cent. reached or exceeded the normal mean, as contrasted with 81 per cent. of a delinquent Hebrew group.

In height and weight the delinquent boys were equal to the norm, with some indication of greater than average physical maturity.

Sensory defects appear to be more frequent among delinquent boys than among normals of the same age. In strength of grip, and in vital capacity, the older delinquent boys were somewhat inferior to the non-delinquent.

In addition to the tests and the questionnaire, certain environmental data were gleaned from the case records of the delinquent boys. Comparative "normal" data were secured by means of a questionnaire filled out anonymously by 3,198 New York City public-school children, selected from different social levels. Comparisons of these data suggest that defective parental supervision and bad home conditions are important factors in juvenile delinquency. Atypical marital status of the parents, *i.e.*, separation, divorce or death, was reported in 45.2 per cent. of the delinquent, and only 19.3 per cent. of the normal groups; over three times as many delinquent as normal children had step-parents; and over seven times as many had been in orphan asylums. Environmental factors such as the mother being employed, size of family, number of rooms at home, showed little or no relation to delinquency when differences in nationality and social status were eliminated.

In another comprehensive study of character and emotional traits in delinquents and incorrigibles, Cady (13) tested several groups of boys twelve and one-half to fourteen and one-half years old, including 150 school boys and seventy boys in a reform school. Five tests were administered: an adaptation of the Woodworth P.D. Sheet; a moral judgment test; and three honesty tests, *viz.*, a self-scoring sentence completion test, a square and circle "coördination" test, and an overstatement test. The boys were all rated for "incorrigibility" by their teachers, and these ratings combined to give a behavior criterion. In order to increase the reliability of the criterion ratings, only those ratings which were based upon a feeling of certainty were included. For these ratings based upon averaged teachers' judgments the reliability was .96. When the scores of the reform school group, the incorrigible group and the corrigible groups were compared, a regular order was obtained upon nearly every test. The corrigible group ranked higher than the incorrigible, and the incorrigible, in turn, higher than the reform school group. The correlations between scores on each of five tests, and the incorrigibility estimates made upon the group of 150 school boys were as follows: squares and

circles, .40; sentence completion, .19; overstatement, .41; moral judgment, .31; and P. D. Sheet, .36. While none of these individual correlations is high, when all five tests were combined into a single battery, the multiple correlation was .58 between the battery and the estimates of incorrigibility. The results of Cady's study indicate that poor emotional control, dishonesty and lack of self-control, as estimated by tests, are definitely related to judgments of incorrigibility based upon overt behavior. Probable incorrigibility could be forecast with fair accuracy from such tests.

Courthial (15), in a recent study of delinquent girls, has attempted to equalize some of the more disturbing environmental factors which make comparisons of delinquents and non-delinquents so difficult. Courthial compared eighty-two delinquent and eighty-two normal girls, ages fourteen to eighteen years, on a number of emotional and character tests. All of the girls in both groups fell within the normal range of intelligence. The girls in the non-delinquent (control) group were paired with the delinquent girls for chronological age, Otis S.A. score, cultural status as determined from the Burdick Apperception Test, and occupational status of father, as rated by the Sims' card (p. 121). Table XII enables us to make a comparison of the differences between the delinquent and the control groups. Each difference has been divided by the S.D. (diff.). When

the  $\frac{D}{S.D. (diff.)}$  is negative, it indicates that the delinquent girls' average was lower than that of the normal group; positive ratios indicate the delinquent girls to be higher.

TABLE XII

A COMPARISON OF EIGHTY-TWO DELINQUENT AND EIGHTY-TWO NORMAL GIRLS UPON A NUMBER OF EMOTIONAL AND CHARACTER TESTS

(from Courthial [15])

Tests	$\frac{D}{S.D. (diff.)}$
1. Pressey X-O (Juvenile form)	
Test I. "things considered wrong" . . . . .	-4 29
Test II. "worries" . . . . .	+3 78
Test III. "interests" . . . . .	+1 44
All three tests . . . . .	+ 13
2. Woodworth-Mathews Questionnaire . . . . .	+7 65
3. Moral Knowledge (seven tests selected from Hartshorne and May series)	-2.18
4. Cheating test: Duplicating, self-scoring technique, used with Army Alpha	+2.65
5. Otis Suggestibility Test . . . . .	+3.30
6. May and Hartshorne Persistence Test	
(Story resistance test) . . . . .	+5.27

Table XII shows that the delinquent girls considered fewer things to be wrong, worried about more things, and tended to have more interests than the non-delinquent. The most striking difference was the greater number of neurotic symptoms reported by delinquent girls upon the Woodworth-Mathews questionnaire. Delinquent girls were more resistant to suggestion, and more persistent than the non-delinquent group. Also, they cheated more, and their knowledge of moral concepts was less exact than the non-delinquent group. On the whole, emotional differences, as between the delinquent and normal girls, were both striking and consistent. This is all the more significant since measured intelligence and background were approximately, at least, equalized. This study shows again the important rôle which character and personality tests may play in the investigation of problems of this kind.

The investigations reported in this section agree in finding juvenile delinquents and incorrigibles not markedly lower in tested intelligence than comparable normal groups. The deviation of the delinquent from the normal is rather in character and personality traits than in abstract intelligence; in poorer emotional stability and self-control and in ignorance of social and moral concepts. Broken homes are also definitely related to juvenile delinquency. Knowledge of these factors has been made more definite and sure through experimental studies with psychological tests.

#### BIBLIOGRAPHY

1. ACHILLES, E. M., "Studies in Recall and Recognition," *Archives Psychology*, 44, 1920.
2. BALDWIN, B. T., *Mental Growth Curve of Normal and Superior Children*, University of Iowa Studies, 2, 1922.
3. BEIK, A. K., "Physiological Age and School Entrance," *Pedagogical Seminary*, 20:277-321, 1913.
4. BONSER, F. G., *The Reasoning Ability of Children of the Fourth, Fifth, and Sixth School Grades*, Teachers College, Columbia University, Contributions to Education, 37, 1910.
5. BOOK, W. F., AND MEADOWS, J. L., "Sex Differences in 5,925 High-school Seniors in Ten Psychological Tests," *Journal Applied Psychology*, 12:56-81, 1928.
6. BRIGHAM, C. C., *A Study of American Intelligence*, Princeton University Press, Princeton, 1923.
7. BRIGHAM, C. C., AND BROLYER, C. R., *Sixth Annual Report of the Com-*

- mission on Scholastic Aptitude Tests*, College Entrance Examination Board, New York, 1931.
8. BRONNER, A. F., *A Comparative Study of the Intelligence of Delinquent Girls*, Teachers College, Columbia University, Contributions to Education, 68, 1914.
  9. BURKS, B. S., "The Relative Influence of Nature and Nurture upon Mental Development," *27th Yearbook, National Society for the Study of Education*, Part I, 219-316, 1928.
  10. BURR, E. T., "Minimum Intellectual Levels of Accomplishment in Industry," *Journal Personnel Research*, 3:207-212, 1924.
  11. BURT, C., *Mental and Scholastic Tests*, London, 1921.
  12. BURT, C., "The Causal Factors of Juvenile Crime," *British Journal Psychology*, 3 (Medical Section): 1-33, 1923.
  13. CADY, VERNON, "The Estimation of Juvenile Incurability," *Journal Delinquency Monographs*, 2, 1923.
  14. CAMERON, A. E., "A Comparative Study of the Mathematical Ability of Boys and Girls in Secondary Schools," *British Journal Psychology*, 16:29-49, 1925.
  15. COURTHIAL, ANDREE, "Emotional Differences of Delinquent and Non-delinquent Girls of Normal Intelligence," *Archives Psychology*, 133, 1931.
  16. COWDERY, K. M., "Measures of General Intelligence as Indices of Success in Trade Learning," *Journal Applied Psychology*, 6:311-330, 1922.
  17. DARSIE, M. L., "The Mental Capacity of American-born Japanese Children," *Comparative Psychology Monographs*, 15, 3, 1926.
  18. DAVENPORT, C. B., AND STEGGERDA, M., *Race Crossing in Jamaica*, Carnegie Institute of Washington Publications, No. 395, 1929.
  19. FAUTH, EMIL, "Testuntersuchungen an Schulkindern nach der Methode des fortlaufenden Addierens," *Archiv für die gesamte Psychologie*, 51:1-20, 1925.
  20. FERGUSON, G. O., "The Psychology of the Negro," *Archives Psychology*, 36, 1916.
  21. FREEMAN, F. N., HOLZINGER, K., *et al.*, "The Influence of Environment on Intelligence, School Achievement, and Conduct of Foster Children," *27th Yearbook, National Society for the Study of Education*, Part I, 103-217, 1928.
  22. GAW, FRANCES, "A Study of Performance Tests," *British Journal Psychology*, 15:374-392, 1925.
  23. GODDARD, H. H., *Feeble-mindedness: Its Causes and Consequences*, The Macmillan Company, New York, 1920.
  24. GOODENOUGH, F. L., "The Reliability and Validity of the Wallin Peg Boards," *Psychological Clinic*, 16:199-215, 1927.
  25. GORING, CHARLES, *The English Convict*, London, 1913.
  26. HEALY, W., *The Individual Delinquent*, Little, Brown and Company, Boston, 1915.

27. HIRSCH, N. D. M., "A Study of Natio-racial Mental Differences," *Genetic Psychology Monographs*, 1:233-406, 1926.
28. HUNTER, W. S., AND SOMMERMEIER, E., "The Relation of Degree of Indian Blood to Score on Otis Intelligence Test," *Journal Comparative Psychology*, 2:257-277, 1922.
29. JAMIESON, E., AND SANDIFORD, P., "The Mental Capacity of Southern Ontario Indians," *Journal Educational Psychology*, 19:536-551, 1928.
30. JONES, H. E., "A First Study of Parent-Child Resemblance," *27th Year-book, National Society for the Study of Education*, Part I, 61-72, 1928.
31. KLINEBERG, OTTO, "An Experimental Study of Speed, and Other Factors in 'Racial' Differences," *Archives Psychology*, 93, 1928.
32. KLINEBERG, OTTO, "A Study of Psychological Differences between 'Racial' and National Groups in Europe," *Archives Psychology*, 132, 1931.
33. LAUTERBACH, C. E., "Studies in Twin Resemblances," *Genetics*, 10: No. 6, 525-568, 1925.
34. LOMBROSO, C., *Criminal Man*, G. P. Putnam's Sons, New York, 1911.
35. MERRILL, M. A., "On the Relation of Intelligence to Achievement in the Case of Mentally Retarded Children," *Comparative Psychology Monographs*, 2, No. 10, 1924.
36. MERRIMAN, CURTIS, "The Intellectual Resemblance of Twins," *Psychological Review Monographs*, 5:33, 1924.
37. MURCHISON, C., *Criminal Intelligence*, Clark University Press, Worcester, Massachusetts, 1926.
38. NORSWORTHY, NAOMI, "The Psychology of Mentally Deficient Children," *Archives Psychology*, 1, 1906.
39. PEARSON, KARL, "On the Laws of Inheritance in Man," *Biometrika*, 3: 131-190, 1904.
40. PETERSON, J., "The Comparative Abilities of White and Negro Children," *Comparative Psychology Monographs*, 1, 1923.
41. PETERSON, J., AND LANIER, L. H., "Studies in the Comparative Abilities of Whites and Negroes," *Mental Measurement Monographs*, 5, 1929.
42. PRESSEY, L. W., "Sex Differences Shown by 2,544 School Children on a Group Scale of Intelligence, with Special Reference to Variability," *Journal Applied Psychology*, 2:323-340, 1918.
43. PRESSEY, S. L., AND TETER, G. F., "A Comparison of Colored and White Children by Means of a Group Scale of Intelligence," *Journal Applied Psychology*, 3:277-282, 1919.
44. SANDIFORD, P., AND KERR, R., "Intelligence of Chinese and Japanese Children," *Journal Educational Psychology*, 17:361-367, 1926.
45. SÉGUIN, E., *Idiocy and Its Treatment by the Physiological Method*, 1866. Reprinted by Teachers College, Columbia University, 1907.
46. SLAWSON, J., *The Delinquent Boy*, Badger, Boston, 1926.



47. SNOW, A. J., "Tests for Chauffeurs," *Industrial Psychology*, 1:30-45, 1926.
48. TALLMAN, G., "A Comparative Study of Identical and Non-identical Twins with Respect to Intelligence Resemblances," *27th Yearbook, National Society for the Study of Education*, Part I, 83-86, 1928.
49. TERMAN, L. M., *et al.*, *The Stanford Revision and Extension of the Binet-Simon Scale for Measuring Intelligence*, Warwick and York, Baltimore, 1917.
50. TERMAN, L. M., *et al.*, *Genetic Studies of Genius*, Stanford University Press, California, vol. I, 1925.
51. THORNDIKE, E. L., "Measurement of Twins," *Archives Philosophy, Psychology, and Scientific Method*, 1, 1905.
52. THORNDIKE, E. L., "The Resemblance of Siblings in Intelligence," *27th Yearbook, National Society for the Study of Education*, Part I, 41-53, 1928.
53. TRABUE, M. R., *Completion Test Language Scales*, Teachers College, Columbia University, Contributions to Education, 77, 1916.
54. WHIPPLE, G. M., "Sex Differences in Army Alpha Scores in the Secondary School," *Journal Educational Research*, 15:269-275, 1927.
55. WILLOUGHBY, R. R., "Family Similarities in Mental-Test Abilities," *27th Yearbook, National Society for the Study of Education*, Part I, 55-59, 1928.
56. WOODWORTH, R. S., "Comparative Psychology of Races," *Psychological Bulletin*, 13:388-396, 1916.
57. YERKES, R. M., "Psychological Examining in the U. S. Army," *Memoirs, National Academy Sciences*, 15, 1921.

## INDEX OF SUBJECTS

(Italics indicate references to Part Two)

- A Scale for Measuring Ascendancy-Submission in Personality (The A-S Reaction Study), Allports, 131
- A Test of Public Opinion (A Survey of Public Opinion on Some Religious and Economic Issues), Watson, 139
- Accomplishment quotient, A.Q., 59-61
- Age scale, description of, 14-32
- Agility tests, Garfields, 53-54
- Aiming tests, 48-50
- Alphabet sorting, 76, 104
- American Council Psychological Examination, Thurstone, 41
- Analogies, tests of, 114-117
- Apperception Test, Burdick, 139
- Apprehension, span of, 68
- Aptitude tests, mechanical, 55-62; music, 168-170; art, 170-171; commercial and vocational, 171-174; professional, 174-177
- Army Alpha Intelligence Examination, description, 32; distribution of scores in general population, 33-34; M.A. and Alpha scores, 34-35; revisions of, 35-36; revised Alpha, 41
- Army Performance Scale, 86
- Art Judgment Test, Moer-Seashore, 170
- Art Test, McAdory, 170
- Art tests, validity of, 178-179
- Arthur's Point Scale of Performance Tests, 86
- Assembling tests, Stenquist, 55-58; Minnesota, 57, 59; Toops', for girls, 58
- Association, tests of, 106-119; free, 106-113; controlled, 113-119
- Astigmatism, 25-27
- Ataxiometer, 45
- Athletic Index, Rogers, 19; calculation of, 20-21
- Attitude questionnaires, 139-141; validity of, 148-149
- Audiometer, 23; use of, 31-34
- Automatograph, 45
- Balancing body, test of, 46
- Behavior Rating Schedules, Haggerty, Olson, Wickman, 118
- Binet's tests, descriptions of, 5-6; revisions of, in America, 7-13; growth curve of, 14-19
- Bookkeeping Test, Elwell-Fowlkes, 171
- Buhler Babytests, 72-73
- Calipers, head, 15; use of, 16
- Cancellation, tests of, 70-75
- Card dealing, 76, 104
- Card sorting, tests of, 60, 75-79; relations to other tests, 81, 84, 102-103, 104-105
- CAVD Intelligence Examination, 118, 36-39
- Cephalic index, 14-16
- Character Education Inquiry tests, 153-162; honesty and trustworthiness, 153-156; cooperation, etc., 157ff.
- Clothing Test, Frear-Coxe, 172
- Code test, description of, 93; results obtained with, 97, 98
- Coefficient of Intellectual Ability, 8
- Colgate Mental Hygiene Tests, Laird, 132-133, 137; Personal Inventory, B2, 132; Personal Inventory, C2, 137
- Color blindness, 28; tests of, 29-30
- Columbia Research Bureau Placement Examinations, 56
- Commercial aptitude tests, 171-174
- Constancy of I.Q., meaning and significance of, 14; theoretical requirements for, 14-17; experimental evidence for, 21-22; factors affecting, 22-24
- Contrasted groups, method of, 125
- Coordination tests, 47-52
- Criminals, psychological tests of, 214-221
- Cylinder test, Witmer, 86-88
- Dart-throwing test, 50
- De Sanctis Scale, 87
- Dearborn form boards, 87
- Dearborn Group Intelligence Tests, Series 1, 92
- Delinquents, tests of, 214-221
- Descriptive or Adjective Rating Scale, 109-110
- Detroit Alpha Intelligence Test, Baker, 39

- Detroit First-Grade Intelligence Test, Engle, 94
- Deuteranopia, 29
- Dexterity, finger, O'Connor's test of, 52; tweezer, O'Connor's test of, 52
- Diagnosis, of failure in school by mental tests, 27-31, 53; in educational problems, 61-62; tests for, 63
- Diagnostic Analysis of Effective Leadership, Personal Inventory, DL, 119
- Diagnostic Test for Introversion-Extroversion, Neymann-Kohlstedt, 137
- Digit-Symbol Substitution Test, 100
- Discrimeter, serial, in Stanford Motor Skills Unit, 54
- Dynamometer, hand, 17; back and legs, 18-19
- Educational achievement, tests of, 54-56; construction of, 57-61; in school work, 61-62; lists of, 62-64
- Educational age, E. A., 59-61
- Educational prognosis tests, 54-55, 63-64
- Educational quotient, E.Q., calculation and use of, 59-61
- Emotional Maturity Scale, Willoughby, 133
- Environment, effects upon general intelligence tests, 23; versus heredity, 185-191
- Ethical Discrimination Test, Kohs, 151
- Ewing tests of visual acuity, 27-28
- Examination in clerical work, Thurstone, 172
- "Experience Variables" Record, Chassell, 133
- Experimental Study of Attitudes toward the Church, Thurstone, 140
- Feeble-minded, tests of, 194-196
- Ferguson Form Boards, 85-86
- Form Board, Minnesota Paper, as test of spatial relations, 81; description and use, 84-86
- Foster children, studies of, 189-191
- Genius, studies of, 191-194
- Gesell's Developmental Schedules, 73-74
- Goddard's revision of the Binet-Simon tests, 7
- Goodenough "Drawing a Man" Scale, 81-82
- Graphic Rating Report on Workers, Scale B, 119
- Graphic rating scales, description of, 108-109; illustrations of, 118-122
- Grip, strength of, 16-18
- Group differences, difficulties in measuring, 196-198; sex differences, 198-201; race differences, 202-211
- Group tests, verbal, 32-54; non-language, 91-99
- Group Will-Temperament Tests, Downey, 152
- Haggerty Intelligence Examination, Delta 1, 94
- Haggerty Intelligence Examination, Delta 2, 40
- Hand test, Thurstone, 80
- Hearing, tests of, 23-25, 30-34
- Height, in body indices, 4-6; relation to intelligence, 6-7; measurement of, 8-11
- Height-weight ratio, 4
- Heredity, studies of, 185-191
- Herring revision of the Binet-Simon tests, 11-12
- Holmgren woolens, in testing color-blindness, 29-30
- Home Economics Test, Engle-Stenquist, 173
- Hyperopia, 25
- Intelligence, general, 3-5; individual tests of, 5-32; group tests of, 32-54; requirements of, 43-45; and performance scales, 87-91, non-language tests of, 91-96; non-verbal and verbal tests of, 96-99
- Intelligence quotient, I.Q., defined, 9; constancy of, 14-17; factors affecting, 22-24; in schools and colleges, 25; distribution of, in the general population, 25-26; value in school work, 27-29; educational guidance and, 29-32; use in group tests, 47-50
- Interest Questionnaire for High School Students, Garretson-Symonds, 141
- Interest questionnaires, illustrations of, 141-144; validity of, 149-150
- International Group Mental Test, Princeton, 94
- Introversion-Extroversion in Terms of Interest, Conklin, 137
- Introversion-Extroversion in Young Children, Marston, 138
- Introversion-Extroversion questionnaires, illustrations of, 137-139; validity of, 148
- Ishihara Test of Color-Blindness, 29-30
- Kent-Rosanoff Free Association Test, 108-113
- Koerth pursuit apparatus, 51-52, 54
- Kohs Block Design Test, 82-83

- Kuhlmann-Anderson Intelligence Tests, 40  
Kuhlmann revision of the Binet tests, derivation of, 10-11; administration of, 13
- Law Aptitude Examination, Ferson-Stoddard, 174
- Learning, tests of, 92-106
- Limen, sensory, 24
- Logical memory, tests of, 126-131
- Man-to-man rating scale, 104-105, 108
- Maze, mental, 96, 98
- Maze tracing, tests of, 101; results with, 103-105
- Measures of Musical Talent, Seashore, 168
- Mechanical ability tests, 55-62, Minnesota, 55, 57-61; MacQuarrie, 61-62
- Mechanical aptitude, tests of, 55-62, Stenquist tests of, 55-58; Detroit examination of, for boys, 61
- Memory, tests of, 119-131; memory span, 119-124; recall memory, 124-126; recognition memory, 126-128; logical memory, 128-131
- Mental age, definition of, 9; meaning of, 14-21; relation to Alpha, 34-35; use in group tests of general intelligence, 47-50
- Mental growth curve, form of, 14-16; of Stanford-Binet, 17-19; when measured in equal units, 19-21
- Mental Hygiene Inventory (House's revision of the Woodworth P. D. Sheet), 134
- Mental maturity, on age scale, 24-25
- Merrill-Palmer Scale of Mental Tests, 74-75
- Michigan pulse-rate test, 22
- Minnesota Paper Form Board Test, 84-86
- Mirror drawing, test of, 102; results with, 103-106
- Morphologic index, definition of, 4; results with, 5-6
- Motor ability scale, Brace, 45-46
- Motor tests, batteries of, 53-54; Garfield's agility tests, 53; Stanford Motor Skills Unit, 54
- Multi-Mental Scale, McCall, 40
- Music Test, Hutchinson and Pressey, 168
- Music tests, list of, 168-170; validity of, 178-179
- Myopia, 25-26
- Nagel Cards, as test of color-blindness, 29-30
- National Intelligence Tests, Scales A and B, 41
- Non-language group tests, description of, 69-70, illustrations of, 91-96; contrasted with verbal tests, 96-99
- North Carolina Rating Scale for Fundamental Traits, Allport, 119
- Numerical or Percentage Rating Scale, 109
- Occupational Intelligence Scale, Barr, 120
- Occupational Interest Blank for Women, Manson, 142
- Opposites tests, 114-117. *See also* Association, tests of.
- Otis Group Intelligence Scale, Adv. Exam., Forms A and B, 42
- Otis Group Intelligence Scale, Primary Exam., 95
- Otis Self-Administering Tests of Mental Ability, 42
- Packing blocks, test of, 60; results with, 81, 84
- Paired associates, tests of, 124-125. *See also* Memory, tests of.
- Peg sorting, 78; Wallin Peg Board Test, 79-80
- Percentiles, as a method in group tests, 46-47
- Perception, span of visual, 65; relations to other tests, 66-69; measurement of, 69-70
- Performance scales, 71-87; compared with verbal tests, 87-91
- Personal Data Sheet, Woodworth, 134
- Personal Traits Rating Scale, Heibredner, 138
- Personality, objective tests of, 151-153; validity and use of, 157-162
- Personality and adjustment questionnaires, list of, 131-137; results with, 146-148
- Personality Inventory, Bernreuter, 135
- Personality Rating Scale, American Council on Education, 120
- Personality Schedule, Thurstones, 135
- Pintner-Cunningham Primary Mental Test, 95
- Pintner Non-Language Mental Test, 95
- Pintner-Paterson Scale of Performance Tests, 76-81
- Point scales, Yerkes-Bridges-Hardwick, 7-8; Herring revision of the Binet tests, 11-12
- Point scores as a method of scoring group tests, 45-56
- Porteus Maze Scale, 83-85
- Pre-school children, tests for, 70-76
- Princeton Universal Scale of Performance Tests, 80

- Professional aptitudes, tests of, 174-177; validity of, 179-183
- Prognosis Test of Teaching Ability, Cox-Orleans, 175
- Prognosis tests, educational, 55, 63-64
- Protanopia, 29
- Psychograph, construction and use of, 135-137
- Pursuitmeter, Miles, 52
- Questionnaires, description of, 122; use and construction of, 123-131; lists of, 131-144; validity of, 144-150
- Racial differences, Negro and white, 202-204; American Indian and white, 204-206; orientals in America, and whites, 206-208; natio-racial differences, 208-211
- Randall's Island Performance Series, 86
- Rating Scale for Teachers, Almy-Sorenson, 120
- Rating scales, description of, 103; varieties of, 104-110; evaluation of rating methods, 110-116; validity of, 116-118; lists of, 118-122
- Rational learning test, 95; modified form of, 95-96
- Reading, and visual perception span, 66, 69
- Recall memory, 124-126; relation to other abilities, 127-128. *See also* Memory, tests of.
- Recognition memory, 126-128. *See also* Memory, tests of.
- Reliability, in intelligence tests, 44; of teachers' marks, 58-59
- Respiratory-height coefficient, 13
- Retained members, method of, 124
- Revision of the Woodworth P. D. Sheet, Mathews, 135
- Rhode Island Intelligence Test, Bird-Craig, 95
- Ring-throwing test, 50
- Rote memory, 119-128. *See also* Memory, tests of.
- Scales test, Griffiths, 46
- Scholastic Aptitude Tests for Medical Schools, Moss-Hunter-Hubbard, 175
- School work, value of I. Q. in, 27-31; relation of general intelligence to, 50-54; educational achievement tests and, 61-62
- Sensory tests, 23-25; visual acuity, 25-28; color-blindness, 28-30; auditory acuity, 30-34
- Sex differences, 198-201
- Siblings, studies of, 185-189
- Sight Singing Test, Hillbrand, 168
- Sigma scores, in psychograph, 135
- Sims' Score Card for Socio-Economic Status, 121
- Snellen chart for testing visual acuity, 25-28
- Social Intelligence Test, Moss-Hunt-Omwake, 152
- Spatial relations test, Minnesota, 81-82, 84
- Specific Interest Inventory, Stewart-Bainard, 142
- Speed rotor, 54
- Spirometer, in determining vital capacity, 13
- Spool packing test, 54
- Stadiometer, in measuring height, 8
- Standardization, importance of, for mental tests, 45
- Stanford Educational Aptitude Tests, Jensen, 176
- Stanford Revision of the Binet-Simon Scale, derivation of, 8-10; administration of, 12-13; growth curve of, 17-21; relation to Alpha, 34-35
- Stanford Scientific Aptitudes Test, Zyeve, 177
- Steadiness, tests of, 43-45
- Stenogauge Test, Bengé, 173
- Stenquist's assembling tests, 55-56
- Stenquist's mechanical aptitude tests, 57-58
- Stilling Plates, use in color-blindness, 29
- Strength tests, 16-22
- Substitution tests of learning, 93-95; results with, 97-100, 104
- Suggestibility Test for Children, M. Otis, 153
- Superior child, studies of, 192-194
- Tachistoscope, use of, 65, 67
- Tapping, tests of, 38-43; relation to other tests, 47, 54
- Terman Group Test of Mental Ability, 42
- Test for Ability to Sell, Moss-Wyle-Loman-Middleton, 174
- Test for Social Attitudes and Interests, Hart, 141
- Test in Fundamental Abilities of Visual Art, Lewerenz, 171
- Test of International Attitudes, Newmann-Kulp-Davidson, 140
- Test of Music Information and Appreciation, Kwalwasser, 169
- Test of Musical Accomplishment, Kwalwasser-Ruch, 169

- Three-hole coordination, test of, 47-48, 49  
 Threshold, sensory, 24  
 Tracing tests, 50-51, 75  
 Twins, studies of, 185-189  
 Types, body, 4; Naccarati's classification, 5; Krietschmer's classification, 6  
 Typewriting Test (Stenographic Proficiency), Blackstone, 174  
  
 Validity, of general intelligence tests, 44;  
     of educational tests, 57-58  
 Vision, tests of, 25-30  
 Vital capacity, measures of, 11-13  
 Vital index, 11-13  
 Vocabulary test, construction of, 117; relation to M. A., 117; importance in general intelligence tests, 28  
 Vocational aptitude tests, list of, 171-174; validity and use of, 179-183  
 Vocational guidance, use of tests in, 54, 61, 77, 82-84; Army Alpha use in, 35-37; general intelligence and other tests in, 211-214  
 Vocational Guidance Test for Engineers, Thurstone, 177  
 Vocational Interest Blank, Strong, 143  
  
 Wallin Peg Boards, 79-80  
 Weight, in body indices, 4-6; correlation with intelligence, 7; measurement of, 9-10  
 Whittier Home Rating Scale, Williams, 121  
 Wiggly Block Test, O'Connor, 82-84  
 Woodworth-Wells association tests, 94, 99; results with, 114-119  
 Worcester Form Board Series, 87  
  
 X-O Tests for Investigating the Emotions, Pressey, 136  
 Yerkes-Bridges-Hardwick Point Scale, 7-8



## INDEX OF NAMES

- Abelson, A. R., 40  
 Abernethy, E. M., 6  
 Achilles, E. M., 125, 127, 200  
 Allison, L. W., 103  
 Allport, F. H., 102, 119, 123, 129, 131, 147  
 Allport, G. W., 123, 129, 131, 147  
 Almy, H. C., 120  
 Anastasi, A., 86, 121-123, 125-127  
 Anderson, R., 40  
 Andrews, B. R., 33  
 Arthur, G., 86  
 Atkinson, W. R., 76-77, 99  
  
 Baker, H. J., 61, 39  
 Baldwin, B. T., 9, 10, 48, 77, 79, 192  
 Barlow, M. C., 77  
 Barr, F. E., 120  
 Barnett, M., 162  
 Beik, A. K., 220  
 Bell, E., 23  
 Benge, E. J., 173  
 Bernreuter, R. G., 126, 135, 147  
 Bickersteth, M. E., 48  
 Bills, M. A., 163  
 Binet, A., 107, 121-122, 3, 5-6, 17  
 Bird, C. E., 95  
 Blackburn, J. M., 49-50  
 Blackstone, E. G., 174  
 Boas, F., 14  
 Bolton, E. B., 125-127, 130-131  
 Bonser, E. G., 115, 200  
 Book, W. F., 30, 200  
 Bovard, J. F., 34  
 Boyce, A. C., 112  
 Brace, D. K., 45, 53  
 Bradshaw, F. F., 111, 115, 117  
 Brainard, P., 142  
 Bregman, E. O., 36, 41  
 Brian, C. R., 50  
 Bridges, J. W., 7, 146  
 Brigham, C. C., 93, 197, 201  
 Bristol, A. S., 101  
 Britten, R. H., 26  
 Brolyer, C. R., 201  
 Bronner, A., 42-43, 80, 88, 94, 123-124, 86, 215  
  
 Brooks, F. D., 98  
 Brown, A. W., 98, 178  
 Brown, W., 71  
 Buhler, C., 72-73  
 Burdick, E., 139  
 Burks, B. S., 24, 28, 189-190  
 Burr, E. T., 212  
 Burt, C., 40, 102-103, 20, 27, 82, 222  
 Burtt, H., 64  
 Buswell, G. T., 69  
  
 Cady, V. M., 126, 146, 219-220  
 Calfee, M., 76, 102-105  
 Cameron, A. E., 198-199  
 Carey, N., 125  
 Carman, A., 16  
 Carothers, F. E., 47-48, 99, 107-108, 115, 121-123, 125-128  
 Carr, H. A., 101  
 Carroll, H. A., 64  
 Carver, D. J., 50  
 Cattell, J. McK., 66  
 Cattell, P., 22  
 Chambers, O. R., 147  
 Chapman, J. C., 61, 167  
 Chassell, J. O., 133-134  
 Chave, E. J., 140  
 Child, E., 45, 75, 79, 104  
 Childs, H. G., 117  
 Clinton, A. J., 104  
 Clothier, R. C., 105  
 Cobb, M. V., 27, 51  
 Collins, S. D., 12, 24, 26  
 Colvin, S. S., 132, 3  
 Conklin, E. S., 137-138, 150  
 Conrad, H. S., 112  
 Cornell, C. B., 79  
 Corning, H. M., 52-53  
 Courthial, A., 220-221  
 Cowdery, K. M., 40, 129-130, 212-213  
 Cox, W. W., 172, 175  
 Cozens, F. W., 34  
 Crafts, L. W., 77  
 Craig, C. E., 95  
 Crockett, A. C., 61



- Cunningham, B., 95  
 Current, W. F., 65  
  
 Dallenbach, K. M., 68  
 Darsie, M. L., 206-207  
 Davenport, C. B., 204  
 Davidson, H., 140  
 Dawley, A., 65  
 Dearborn, W. F., 69, 93, 23-24, 87, 92  
 DeBusk, B. W., 12  
 DelManzo, M. C., 152, 158-159  
 DeSanctis, S., 87, 100  
 DeWeerd, E. H., 73  
 Dewey, E., 45, 75, 79, 104  
 Dickson, V., 65  
 Dixon, R. B., 14, 27  
 Dodd, S. C., 94, 99  
 Downey, J. E., 152, 159  
 Dreyer, G., 13  
 Dunlap, K., 41  
  
 Earle, F. M., 100  
 Edgerton, H., 52  
 Edwards, D. J., 13  
 Elliot, R. M., 40, 46, 49, 58, 38  
 Elwell, F. H., 171  
 Engel, A. M., 93  
 Engle, E. M., 173  
 Ewing, A. E., 27  
  
 Fauth, Emil, 199  
 Ferguson, G. O., 85, 88, 202  
 Fernald, M. R., 64  
 Ferson, M. L., 174-175, 181  
 Fischer, C. R., 16  
 Fisher, V. E., 98  
 Flemming, E. G., 136, 146, 150  
 Fletcher, H., 31  
 Foster, J. C., 8  
 Foster, W. S., 66, 68, 93  
 Fowler, E. P., 34  
 Fowlkes, J. G., 171  
 Fox, E., 23  
 Franz, S. L., 62, 145  
 Frear, F. D., 172  
 Freeman, F. N., 24, 100, 190-191  
 Freyd, M., 109, 123-125, 174  
  
 Galton, F., 6, 103  
 Garfiel, E., 17, 40, 43-44, 47, 50, 53-54  
 Garretson, O. K., 128-129, 141, 149  
 Garrett, H. E., 71, 98-99, 121-122, 126, 130, 65, 159  
 Garrison, K. C., 76, 98  
 Gates, A. I., 7, 71, 77, 122, 125, 130, 49, 51  
 Gates, G. S., 41  
  
 Gaw, F., 80, 89, 199  
 Gesell, A., 23, 73-74  
 Gittings, I. E., 7, 12  
 Goddard, H. H., 8, 12, 7  
 Goodenough, F. L., 50, 79-80, 122, 65, 81, 200  
 Goring, C., 214  
 Graves, K., 23  
 Gray, C. T., 66, 69  
 Griffiths, C. H., 43-44, 46, 48-49  
 Guilford, J. P., 108  
 Guinzberg, R. L., 7  
 Gullette, R., 111  
  
 Haggerty, M. E., 40-41, 94, 98, 118  
 Hallowell, D. K., 80, 120  
 Hanson, G. F., 13  
 Hardwick, R. S., 7-8  
 Harris, D., 112  
 Hart, H. N., 127, 141  
 Hartmann, G. W., 123  
 Hartshorne, H., 113, 115, 117, 140, 151, 153, 156, 160-162  
 Haught, B. F., 95, 98  
 Haupt, I. A., 29  
 Hayes, M. H., 65  
 Healy, W., 42-43, 80, 88, 94, 123-124, 86, 222  
 Heidbreder, E., 123-124, 138-139  
 Herring, J. P., 7, 11-12, 100  
 Hertzberg, O. E., 46, 70  
 Hildreth, G., 21  
 Hillbrand, E. K., 168  
 Hirsch, N. D. M., 208-209  
 Hoffman, J., 68  
 Hoitsma, R. K., 148, 150  
 Hollingworth, H. L., 39, 43, 47, 116, 145  
 Hollingworth, L. S., 17  
 Holzinger, K. J., 100, 125  
 House, S. D., 123, 126, 134  
 Howe, E. C., 12  
 Hubbard, H. F., 175  
 Hubbard, R. M., 125  
 Hull, C. L., 15, 39, 43, 66, 167, 183  
 Hunt, T., 152-153, 159, 160  
 Hunter, O. B., 175  
 Hunter, W. S., 204  
 Huntington, E., 14  
 Hurlock, E. B., 120-121, 129  
 Husband, R. W., 105  
 Hutchinson, H. E., 168  
  
 Imm, A. J., 78  
 Itard, J., 6  
  
 Jacobson, C., 111

- Jamieson, E., 205  
 Jarman, B. L., 26  
 Jastrow, J., 108  
 Jensen, M. B., 176, 180  
 Johnson, B. J., 17, 45, 50, 87, 80, 87, 89  
 Jones, H. E., 78, 25, 188-189  
 Jung, C. J., 111  
  
 Kcal, H. M., 66  
 Kelauber, G. N., 57-58, 50  
 Kelley, T. L., 5, 27, 61  
 Kellogg, W. N., 159  
 Kempl, G. A., 24, 26  
 Kent, G. H., 82, 108-113, 101  
 Kerl, R., 207  
 Kincaid, M., 73  
 King, H. B., 101  
 Kitson, H. D., 40  
 Klineberg, O., 24, 198, 202, 206, 209-210  
 Knollen, H. E., 18  
 Knotts, J. R., 101  
 Koerth, W., 51-52, 54  
 Kohlstedt, K. D., 137  
 Kohls, S. C., 82, 87, 89, 151-152, 160  
 Kornhauser, A. W., 115  
 Kraepelin, E., 6  
 Krietschmer, E., 5-6  
 Kroeber, A. L., 14  
 Kuderna, J. G., 36  
 Kuhlmann, F., 7, 10, 13, 22, 24, 11  
 Kulp, D. H., 140  
 Kwalwasser, J., 169  
  
 Laird, D. A., 119, 132-133, 137, 145-146  
 Landis, C., 111  
 Lanier, L. H., 41, 97-99  
 Laurer, F. A., 35  
 Lauterbach, C. E., 186-187  
 Leach, H. M., 133  
 Lemmon, V. W., 71, 99, 131  
 Levy, H. H., 26  
 Lewerenz, A. S., 171, 179  
 Lincoln, E. A., 93, 100  
 Link, H. C., 39-40, 77, 81  
 Loman, W., 174  
 Lombroso, C., 214  
 Lowe, G., 62, 99  
 Lowell, F., 112  
  
 McAdory, M., 170-171  
 McCall, W. A., 73, 28, 40  
 McGeech, J. A., 125, 164  
 MacPhail, A., 51  
 MacQuarrie, T. W., 61-62  
 Maller, J., 117  
  
 Manchester, G. S., 108  
 Manson, G. E., 142  
 Marston, L. R., 126, 138  
 Martin, A. L., 101  
 Mathews, E., 135-136, 146  
 May, M., 113, 115, 117, 140-151, 153, 156, 160-162  
 Mead, M., 23  
 Meadows, J. L., 200  
 Meier, N. C., 158, 170, 179  
 Merrill, M. A., 223  
 Merriman, C., 186  
 Middleton, W., 174  
 Miles, W. R., 45, 52, 54, 101  
 Milner, M., 100  
 Monahan, J. E., 17  
 Morgenthau, D. R., 83  
 Moss, F. A., 152, 174-176, 181  
 Mulligan, J. H., 15  
 Munsterberg, H., 112  
 Murchison, C., 215-217  
 Murdoch, K., 7, 15  
 Murphy, G., 110-111  
 Muscio, B., 41  
  
 Naccarati, S., 4-7, 12  
 Newman, G. B., 140  
 Newmark, E. D., 120-121, 129  
 Neymann, C. A., 137  
 Nöh, E. J., 108  
 Norsworthy, N., 122, 194  
 Nyswander, D. B., 105  
  
 O'Brien, J. A., 66, 69  
 O'Connor, J., 52, 82-84  
 Olson, W. C., 118  
 Omwake, K. T., 152  
 Orleans, J. S., 175  
 O'Rourke, L. J., 159  
 Otis, A. S., 42, 95  
 Otis, M., 112, 153  
  
 Paschal, F. C., 86  
 Paterson, D. G., 14-15, 40, 46, 49, 57-58, 88, 101, 111, 113, 115, 119  
 Pearson, K., 6, 14-15, 103, 187, 189  
 Peatman, J. G., 5  
 Peterson, H. A., 36  
 Peterson, J., 77, 95-99, 103, 203  
 Pinel, P., 6  
 Pintner, R., 88, 24, 51, 95, 101  
 Popenoe, H., 60  
 Porteus, S. D., 134, 84, 87, 89  
 Postle, D. K., 52  
 Poull, L. E., 100  
 Pressey, L. W., 168, 200

- Pressey, S. L., 126, 136-137, 147, 203  
 Proctor, W. M., 27, 37  
 Pyle, W. H., 36, 77-79, 100, 104, 107, 115-116, 120-121, 129-130  
 Radosavljevich, P. R., 14  
 Rand, G., 49, 101  
 Ranschburg, P., 68  
 Raubenheimer, A. S., 151  
 Ream, M. J., 63, 125  
 Reid, R. W., 15  
 Renshaw, S., 52  
 Rogers, F. R., 4, 12-13, 19, 21-22  
 Rogers, M. C., 23  
 Rosanoff, A. J., 108-113  
 Rosanoff, I. R., 112  
 Ross, E. L., 90  
 Ruch, G. M., 76-77, 93, 97, 58, 152, 158-159, 169  
 Rudisill, E. S., 17  
 Rugg, H. O., 114  
 Ruml, B., 45, 75, 79, 104  
 Sandiford, P., 26, 205, 207  
 Sangren, P. V., 98  
 Schiefflin, B., 88  
 Schneek, M. R., 115, 117, 143  
 Schriefer, L., 87, 80, 89  
 Schuyten, M. C., 16  
 Schwesinger, G. C., 88  
 Scott, W. D., 104-105  
 Seashore, C. E., 31-32, 170, 179  
 Seashore, R. H., 40, 53-54  
 Seashore, S., 52, 54  
 Seguin, E., 6, 196  
 Shakow, D., 101  
 Shaw, E. A., 100  
 Sheldon, W. H., 5-6  
 Shen, E., 115  
 Sherman, E. B., 15  
 Shimberg, M., 62, 99  
 Simon, T., 5  
 Sims, V. M., 121  
 Slawson, J., 146, 217-219  
 Smedley, F. W., 13, 16, 18, 42  
 Smith, H. L., 9-10, 67  
 Smith, W. W., 111  
 Snadden, G. H., 77  
 Snoddy, G. S., 102  
 Snow, A. J., 213-214  
 Sommermeier, E., 204  
 Sommerville, R. C., 6, 15, 40, 44, 71  
 Sorenson, H., 120  
 Spearman, C., 114, 115  
 Spence, K. W., 103  
 Squires, P. C., 80  
 Stalnaker, E. M., 18  
 Starch, D., 102, 104, 58  
 Starr, A. S., 124  
 Stecher, L. I., 48, 77, 79  
 Steggerda, M., 204  
 Stenquist, J. L., 55-58, 173  
 Sterling, E. B., 23  
 Stern, W., 3  
 Stewart, F. J., 142  
 Stockard, C. R., 16  
 Stoddard, G. D., 67, 174-175, 181  
 Strong, E. K., 123, 129-131, 148-149  
 Stutsman, R., 79-80, 88, 75  
 Sullivan, L. R., 7, 15  
 Swanson, C. A., 78  
 Symonds, P. M., 112, 116-117, 127, 141  
 Tallman, G., 187  
 Taylor, J. F., 27  
 Teagarden, F. M., 25  
 Terman, L. M., 106-107, 117, 120-122, 3, 4, 7, 8-9, 12, 18, 21, 24, 27, 28-29, 41-42, 52, 120, 146, 192, 199-200  
 Terman, S. W., 37  
 Teter, G. F., 203  
 Thorndike, E. L., 118, 4, 20, 25, 28, 41, 43, 101, 118, 186, 187-188  
 Thurstone, L. L., 5, 80, 94, 19, 20, 41, 51, 124, 126, 131, 135, 140, 144, 146, 150, 172, 177, 180, 181  
 Thurstone, T. G., 41, 124, 126, 135, 146  
 Tinker, M. A., 68, 78, 93  
 Titchener, E. B., 64  
 Toops, H. A., 58, 167  
 Townsend, S., 103  
 Trabue, M. R., 33, 200  
 Tredgold, A. F., 37  
 Turner, A. H., 12-13  
 Turner, E. M., 90  
 Uhrbrock, R. S., 159  
 Vickery, K., 71  
 Voelker, P. F., 151  
 Wallin, J. E. W., 79, 68  
 Warden, C. J., 103  
 Washburn, M. F., 133  
 Watson, G. B., 123, 127, 139, 148  
 Weidensall, J., 104  
 Weiss, A. P., 52  
 Wellman, B., 50-51, 75  
 Wells, F. L., 52, 73, 114-115, 118-119, 36  
 Wheat, L. B., 112  
 Whipple, G. M., 4, 7, 12-13, 16, 18, 33, 42, 49-50, 66-69, 88, 129, 41, 199

- Whitely, P. C., *164*  
Whitley, M. T., 88, 118  
Whitman, E. C., 52  
Wickman, E. F., *118*  
Williams, J. F., 13  
Williams, J. H., *121*  
Willoughby, R. R., 97, *133, 188-189*  
Wilson, M. G., 13  
Wissler, C., 16, 122  
Witmer, L., 86-87  
Witty, P. A., 27  
Wood, B. D., *51, 58*  
Wood, T. D., 9-11  
Woodrow, H., 112  
Woodworth, R. S., 73, 114-115, 118-  
119, 3-4, *123, 134-135, 145-146*  
Woolley, H. T., 16, 79, 74  
Worthington, M. R., 90  
Wright, W. W., 9-10, 67  
Wyatt, S., 115  
Wyle, H., *174*  
Wylie, A. T., 115-116  
Wyman, J. B., *149-150*  
  
Yerkes, R. M., 91, 7-8, *41, 101*  
Yoakum, C. C., 91, 105, *68, 101*  
  
Zyve, D. S., *177, 182*
-

















